The Biometric Society

BIOMETRICS

FOUNDED BY THE BIOMETRICS SECTION OF THE AMERICAN STATISTICAL ASSOCIATION

TABLE OF CONTENTS

Number 2 June 1953 Volume 9

Material for *Biometrics* should be addressed to Miss Gertrude Cox, Institute of Statistics, Box 5457, Raleigh, North Carolina, except that authors residing in one of the following organized regions can expedite the handling of their papers by submitting them to the Assistant Editor for that region.

British Region: Dr. M. J. R. Healy, Rothamsted Exp. Sta., Harpenden, Herts, England (serving in Dr. Finney's absence); **Australasian Region:** Dr. E. A. Cornish, University of Adelaide, Adelaide, Australia; **French Region:** Dr. Georges Teissier, Faculte des Sciences de Paris, 1 rue V. Cousin, Paris, France

Material for Queries should go to Professor G. W. Snedecor, Statistical Laboratory, Iowa State College, Ames, Iowa.

Articles to be considered for publication should be submitted in triplicate.

The Biometric Society is an international society devoted to the mathematical and statistical aspects of biology and welcomes to membership biologists, mathematicians, statisticians and others who are interested in its objectives. Through its regional organizations the Society sponsors regional and local meetings. National secretaries serve the interest of members in Italy, Denmark, the Netherlands, India, Germany and Japan and there are many members "at large". Dues in the Society for 1953 for residents of the Western Hemisphere are as follows: Full membership including subscription to *Biometrics* is $7.00. Members of the Biometrics Section of the American Statistical Association who subscribe to the journal through that organization may become members of The Biometric Society on the payment of $3.00 annual dues. For members in other parts of the world, full membership including subscription to *Biometrics* is $4.50, except that members who subscribe to the journal through the American Statistical Association pay annual dues of $1.75. Information concerning the Society can be obtained from the Secretary, The Biometric Society, Drawer 1106, New Haven 4, Connecticut, U.S.A.

Annual subscription rates to non-members are as follows: For American Statistical Association Members, $4.00; for subscribers, non-members of either American Statistical Association or The Biometric Society, $7.00. Subscriptions should be sent to the Managing Editor, *Biometrics*, P. O. Box 5457, Raleigh, North Carolina, U.S.A.

Entered as second-class matter at the Post Office at New Haven, Conn., under the Act of March 3, 1879. Additional entry at Richmond, Va. Business Office, 52 Hillhouse Ave., New Haven, Conn. *Biometrics* is published quarterly—in March, June, September and December.

THE ANALYSIS OF VARIATION IN A NATURAL POPULATION OF LADY BEETLES

FRANK M. STURTEVANT, JR., LYLE D. CALVIN AND ORLANDO PARK

From the Division of Biological Research, G. D. Searle & Co., Box 5110, Chicago 80, Institute of Statistics, North Carolina State College, Raleigh; and Cresap Biological Laboratory, Northwestern University, Evanston

INTRODUCTION

Biometric and taxonomic studies of quantitative characters in natural populations have led to the discovery of cases of continuous geographic variation, or clines, and of subpopulations displaying various degrees of reproductive and morphological intergradation. Because of the empirical fact that two or more natural entities may bear *any* degree of relationship along the spectrum of reproductive isolation, from complete intersterility to complete interfertility, regardless of their absolute or partial geographic coexistence, only an arbitrary point can be established on that spectrum for the taxonomic convenience of division of such entities into subgroups. The polytypic species concept (Mayr, 1942, Chap. 6) is a function of great subjectivity when applied to geographically separated, and especially morphologically similar, populations. The danger lies in the necessity of classifying microevolutionary units for ease of handling and in the implicit assumption by many taxonomists that such classifications reflect the true biological order of nature. Many authors have vainly sought a species definition for lack of realization of this latter point (see discussion by Gilmour, 1951).

One of the most objective methods for investigating the systematic relationships between evolutionary units is that of Womble (1951), even though some subjectivity may enter the picture in the weighting of various characters. Fisher (1936) has dealt with such weightings by the use of discriminant functions. In order to apply Womble's differential systematics, a good deal of information first must be accumulated on the distribution of points on the geno-, pheno-, or ecocline. The genocline is the most fundamental of the three and much work

has been pursued on geographic variation in the frequencies of certain genes.

A prime requisite for genocline analysis is the characterization of the genetics and variation in a natural population. Shull (1944) has published the results of genetic studies on several populations of the lady beetle *Hippodamia convergens* Guer. This species of coccinellids is one of the most common of the lady beetles and is found in abundance throughout North America. One of its distinguishing characteristics is the presence on each elytron, or wing-cover, of six black spots or maculations. Mainly through Shull's efforts, the heredity of the elytral maculations in this species and the genetics of interspecies relationships have been well-developed (Shull, 1949). A review of the literature on the genetics of the Coccinellidae appeared in an earlier paper by Shull (1943).

The purpose of the present paper is to investigate the variation of a genetic character in a natural population of *Hippodamia convergens*, illustrating a method for the separation of overlapping distributions of phenotypes.

THE SPOTLESS PHASE

The elytral maculations of *Hippodamia convergens* vary greatly in size and may be wholly or partly absent, in which case the condition is known as the spotless phase. Since the phenotypes of the spotless and spotted phases overlap, Shull (1946) devised a method for their separation. After many genetic crosses, he found that spotless beetles could be identified best by considering only the posterior three maculations, which, he observed, tended to be larger than the anterior three in spotted beetles and to show a greater frequency of absence in spotless beetles. An arbitrary unit for the measurement of the greatest diameter of a spot was established such that no maculation was scored as greater than four; therefore, no total of the three posterior spots could exceed $3 \times 4 = 12$.

In Shull's laboratory population, which had been derived from a single pair of Michigan beetles (Shull, 1951), it was found that the inheritance of spot diameters was such that a posterior total of 7.74 or less would divide, on the average, the spotted and spotless phases (Shull, 1946). It was further established that the spotless phase was governed by "one almost dominant gene", plus several modifiers (Shull, 1944).

Since Shull, then, had established that the size of the elytral maculations in this species was under genetic control, it was first necessary to examine the variation of the posterior total relative to the anterior

total, and second to establish a method for separating the spotless and spotted components of the natural population at hand. It was not to be expected, on *a priori* grounds, that the area of overlap would be the same for a highly inbred laboratory stock and a natural population, because of the genetic drift which so often accompanies inbreeding.

MATERIAL AND METHODS

A hibernating population of 1,021 *Hippodamia convergens* beetles was collected in a clump of *Calamovilfa* grass in the foredune stage of the Ogden Dunes, Porter County, Indiana on October 26, 1946, at 11:00 AM. The right elytra from 1,011 of these specimens were removed and affixed to slides, ten elytra being damaged in the preparation. The greatest diameter of each maculation was then measured microscopically in units ranging from 0 to 8.

In accordance with Shull's observations (1944) that the three posterior maculations appeared to act as a unit, totals of the greatest diameters of these spots for each elytron were made and their frequencies tabulated (table 1). The same was then done for the three anterior spots on 124 beetles chosen at random. The regression line of the anterior sums (y) on the posterior sums (x) for these 124 elytra was calculated by least squares: $y = 2.17 + 0.344x$ (fig. 1). The standard deviation of the slope was $\pm$ 0.108. The coefficient of correlation was 0.67, with 99 per cent fiducial limits at 0.78 and 0.52.

TABLE 1.

Frequencies of the totals of the three posterior maculations of the right elytra of 1,011 *Hippodamia convergens*

Class	Frequency	Class	Frequency
0	39	11	95
1	10	12	123
2	8	13	147
3	9	14	131
4	9	15	106
5	13	16	59
6	15	17	59
7	23	18	15
8	26	19	8
9	37	20	7
10	71	21	0
		22	1

It was evident that the posterior spots, as a group, tended to be larger than the anterior in the mixed sample (fig. 1). The positive y-intercept does not mean that *individually* the posterior three spots were most frequently absent. The frequency of absence of each spot, numbered anterior to posterior, was calculated and entered in table 2. Similarly, of the total occurrence of spot-absence of 381 cases, the individual spots accounted for the percentages listed in the last column of table 2. The posterior spots were absent more frequently than the anterior. These two findings agree with Shull's observations cited earlier.

With the exception of spot I, the posterior three maculations showed a greater frequency of absence and accounted for greater fractions of the total spot-absence in the mixed sample. Since, as a group, the posterior spots behaved as predicted by Shull's observations, although spot I also exhibited a high tendency to be absent, the following calculations were based on the posterior sums. Such procedure would enable direct comparison of the present results with those of Shull.

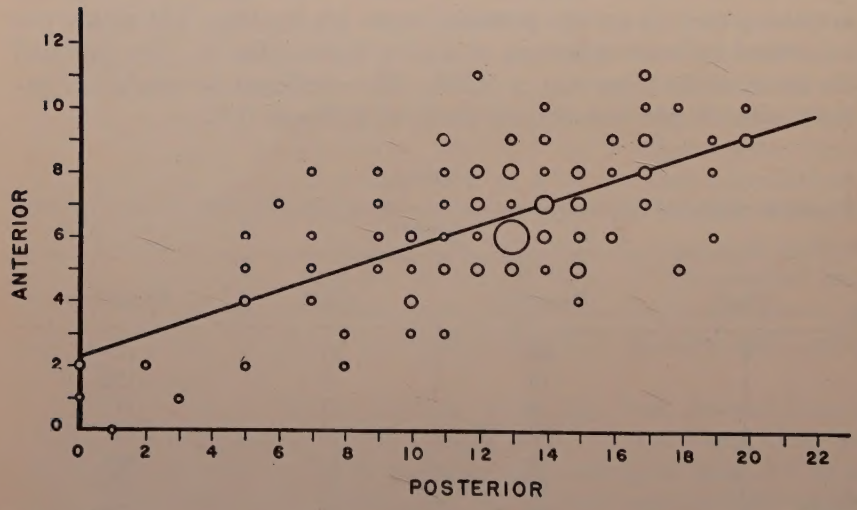

FIG. 1.

Regression of the greatest-diameter sums of the three anterior on the three posterior maculations of 124 right elytra selected at random from the natural population of 1,011 *Hippodamia convergens* beetles. The number of elytra at any one point varies from one to eleven, and is represented by different sized circles.

TABLE 2.

Absences of each elytral maculation, numbered anterior to posterior, in 124 *Hippodamia convergens* chosen at random from the natural population

Spot	Cases of Absence		
	Number	Per cent of possible	Per cent of all spots absent
I	98	79	26
II	28	23	7
III	35	28	9
Anterior	161	43	42
IV	52	42	14
V	55	44	14
VI	113	91	30
Posterior	220	59	58
Totals	381		100

ESTIMATION OF POPULATION PARAMETERS

Employing the data in table 1, a histogram was constructed from the frequencies of the sums of the greatest diameters of the three posterior maculations on 1,011 right elytra (fig. 2). Since the units of measurement used in this paper are twice those employed by Shull, the point on the abscissa which would separate the spotless and spotted beetles according to his method would be twice his separation point, that is, $2 \times 7.75 = 15.50$. From inspection of figure 2, it is seen that this value is much greater than it should be for the present population. There was only a slight probability that the separation points of the two populations would coincide, because the population at hand was a natural one, while Shull's was a laboratory stock derived from a single pair of beetles. Such a cultured stock, especially when small, is highly susceptible to the effects of genetic drift, and as a result, one would expect *a priori* a shift in the separation point and mean of the spotted beetles. The standard deviation of the spotted population was probably smaller in Shull's stock, since he had a separation point at 15.50, when the maximal possible value was 24.00 (in our units). Obviously, some method for separating the spotted and spotless phases other than Shull's would have to be utilized in the present case.

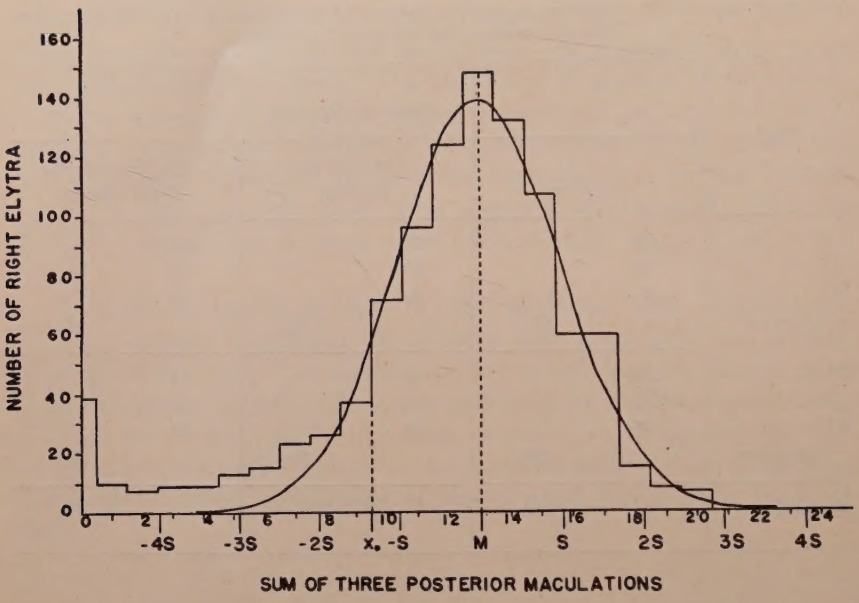

FIG. 2.

Histogram of frequencies of greatest-diameter sums of the three posterior maculations of 1,011 right elytra from a natural population of *Hippodamia convergens*. Mean (M) and standard deviation (S) calculated by truncating the histogram at x_0 ; the corresponding normal curve is imposed on the diagram. Area under the normal curve represents the 897 spotted beetles; remaining area to the left represents the 114 spotless beetles.

The histogram in figure 2 is obviously bimodal, and the population to the right (the spotted beetles) seemed to describe a normal curve. It was reasoned therefore that the two populations could be separated best by cutting the curve at an arbitrary truncation point (x_0), beyond which it would be highly improbable that any spotless beetles would fall, and then calculating the mean and standard deviation of the truncated normal curve. From these statistics, the size of the population under the truncated curve could be calculated. The estimation of the total numbers of spotted and spotless beetles from these data would then be a relatively simple step. The method used for working with such a truncated normal curve with unknown population size was that of Cohen (1949). His notation has been followed in the present paper with the exception that the primes have been omitted.

Fitting the Normal Curve

The point of truncation was selected by inspection as $x_0 = 9.5$. This point is an arbitrary choice and might have been taken slightly lower. It was taken as the point above which, in the authors' opinion, there was a very low probability that any spotless beetles would occur.

The necessary summations are $\sum x = 3288.0$ and $\sum x^2 = 17{,}205.5$, where x is measured from x_0 . The expression

$$\psi_1 = \frac{n \sum x^2 - (\sum x)^2}{(\sum x)^2}$$

was calculated as 0.30820, where n is the number of beetles (822) in the truncated sample. ψ_1 is needed only in estimating h and σ, and is a moment function of h (the point of truncation measured in standard units; that is, $h = (x_0 - \mu)/\sigma$, where μ and σ are the mean and standard deviation respectively of the population).

Cohen has plotted the relationship between ψ_1 and h, so that an initial estimate of $h = -1.40$ could be obtained from the graph. By successive approximations in the expression

$$\psi_1 = \left(\frac{1}{Z - h}\right)\left(\frac{1}{Z - h} - Z\right),$$

h was calculated as -1.383 and Z as 0.16725. Z is defined by the identity

$$Z(h) \equiv \phi(h)/I_0(h),$$

where $\phi(h)$ is the ordinate of the normal curve at $t = h$, and $I_0(h)$ is the area under the curve to the right of the point $t = h$.

The estimates of μ and σ are given by

$$m = x_0 - hs$$

and

$$s = \frac{1}{n} \cdot \frac{\sum x}{(Z - h)}.$$

To obtain estimates of the standard deviations of m and s, ten independent random samples of 1011 elytra each were drawn from the total population of 1011 and separate estimates of m and s calculated for each sample. The standard deviations of m and s, each with 9 degrees of freedom, were obtained from these 10 samples. The values

of m and s, with their standard deviations, are

$$m = 13.068 \pm 0.179$$

$$s = 2.580 \pm 0.081$$

Having calculated the mean and standard deviation, the truncation point in standard units was given by $h = -1.383$. The area to the right of x_0 is calculated by

$$\int_{-1.383}^{\infty} f(t)\,dt = \int_{-1.383}^{0} f(t)\,dt + \int_{0}^{\infty} f(t)\,dt,$$

where

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \qquad \text{(Kenney, 1947, p. 116 ff.).}$$

From the table for the normal curve,

$$\int_{-1.383}^{\infty} f(t)\,dt = 0.4166 + 0.5000 = 0.9166.$$

The total number of beetles in the spotted population is calculated from $0.9166\,N = 822$. $N = 896.8$ or 897 and the number of beetles in the spotless population is $1011 - 897 = 114$. The standard error of N is estimated by replicate calculations of N from the 10 random samples of 1011 elytra each; $s_N = 14$.

To impose the normal curve on the histogram in figure 2, values of x were substituted in the equation $t = (x - m)/s$ and the corresponding standard-unit ordinates read from the table for the normal curve. These were converted in turn by substitution in the expression $y = N \cdot f(t)/s$, where $f(t)$ is the ordinate (Kenney, 1947, p. 125). The solutions for y are plotted in figure 2 against the appropriate values of x.

The goodness of fit of the experimental data to the right of the point of truncation was tested by $\chi^2 = 13.02$, d.f. $= 9$, $P = 0.17$.

At the suggestion of Wright (1952), a semilog parabola was fitted by least squares to the data after plotting the log frequencies in table 1 against the classes. The goodness of fit was tested as before: $\chi^2 = 25.94$, d.f. $= 10$, $P < 0.01$. Obviously, the normal curve proved a better fit to the data than the semilog parabola.*

The Distribution of Spotless Beetles

After removing the spotted beetles from the mixed population by either of the above two methods, the remaining spotless beetles assumed

*Dr. G. E. P. Box has pointed out that, if the log frequencies are suitably weighted in a least squares analysis, this procedure should give improved estimates of the mean and variance which would yield a smaller χ^2 for the goodness of fit test.

a uniform distribution, except for a mode at class zero (fig. 2). This mode could have been produced by the necessary truncation at this point. In contrast to Shull's evidence (1944) that his laboratory population contained some four modifiers of the gene *Spotless*, Wright (1952) commented that the present distribution of spotless beetles would not be expected if variability depended on "some four more or less equally frequent and equally important modifiers. There is some suggestion of one main modifier or, perhaps more probably, an indication of such inequality in the physiological significance of scale units in this region that no interpretation is warranted."

The Question of Gene Frequencies

Shull (1944) applied his criterion of the spotless phase to several widely distributed populations of *Hippodamia convergens*. Using the Hardy-Weinberg equilibrium formula*, which disregards systematic, random and nonrecurrent changes in gene frequency (Wright, 1949), the frequencies of the gene *Spotless* for each population were calculated. Further, assuming four partially dominant, equally abundant and effective modifiers of *Spotless*, the frequencies of the modifiers were calculated**.

Similar computations with the present data yielded a frequency of *Spotless* of $1 - (897/1011)^{\frac{1}{4}} = 5.81$ per cent, which compares favorably with Shull's values of 5.84 and 6.45 for two Michigan populations. However, such statistics are only rough approximations, because of the large phenotypic overlap, the disregard of such factors as mutation and selection pressures, and the indirect evidence relating to the inheritance of the spotless phase in nature. As was noted earlier, genetic drift is common in small laboratory stocks and, therefore, extrapolation to natural populations cannot be made with a high degree of confidence.

On the other hand, genetic drift would not be expected *a priori* in natural populations of *Hippodamia*, according to the known ecology of the genus (Cutright, 1924; Park, 1930; Balduf, 1935). Although hibernation occurs in small restricted groups (Allee *et al.*, 1949, p. 538), mating does not occur until these groups have dispersed in the spring; thus genetic drift would be held to a minimum and neighboring populations would not be expected to differ greatly in gene frequencies.

---

*$\left[\sum_{i=1}^{k} (q_i S_i)\right]^2$, where q_i is the gene frequency of the spotless gene S_1.

**$1 - (P_0/P_s)^{1/8}$, where P_0 is the number of beetles of class zero (table 1) and P_s is the total number of spotless beetles.

SUMMARY

A natural population of the lady beetle *Hippodamia convergens* Guer. was submitted to a statistical analysis of its elytral pattern in order to separate two overlapping phenotypes. These phenotypes were the so-called "spotless" and "spotted" phases, the inheritance of which had been investigated earlier by A. F. Shull. It was shown that the three posterior elytral spots, as a group, tended to be absent more frequently than the anterior three, and also tended to be larger in size. A normal curve was fitted to the arbitrarily truncated frequency distribution of the total of the greatest diameters of the three posterior maculations of spotted beetles. The normal curve gave an acceptable fit as tested by chi-square. From the statistics computed, the population sizes of the spotted and spotless beetles were calculated. The question of gene frequencies in this and other populations was discussed.

REFERENCES

Allee, W., A. Emerson, O. Park, T. Park, and K. Schmidt. *Principles of animal ecology.* Saunders, Phil., 1949.

Balduf, W. *The bionomics of entomophagous Coleoptera.* John Swift, Chicago, 1935.

Cohen, A. On estimating the mean and standard deviation of truncated normal distributions. *J. Amer. Stat. Assn. 44:* 518, 1949.

Cutright, C. Bionomics of *Hippodamia tridecem-punctata L. Ann. Ent. Soc. Amer. 17:* 188, 1924.

Fisher, R. The use of multiple measurements in taxonomic problems. *Ann. Eug. 7:* 179, 1936.

Gilmour, J. The development of taxonomic theory since 1851. *Nature 168:* 400, 1951.

Kenney, J. *Mathematics of statistics, part one.* Van Nostrand, N.Y., 1947.

Mayr, E. *Systematics and the origin of species.* Columbia Univ. Press, N.Y., 1942.

Park, O. Studies in the ecology of forest Coleoptera. *Ann. Ent. Soc. Amer. 23:* 57, 1930.

Shull, A. Inheritance in ladybird beetles. I. The spotless and spotted elytra of *Hippodamia sinuata* L. *J. Hered. 34:* 329, 1943.

——. Inheritance in ladybird beetles. II. The spotless pattern and its modifiers in *Hippodamia convergens* and their frequency in several populations. *J. Hered. 35:* 329, 1944.

——. The standards by which the spotless phase of *Hippodamia convergens* is judged. *Ann. Ent. Soc. Amer. 39:* 190, 1946.

——. Extent of genetic differences between species of *Hippodamia* (Coccinellidae). *Proc. 8th Int. Cong. Gen., Hereditas suppl. vol.:* 417, 1949.

——. Personal communication. Sept. 22, 1951.

Womble, W. Differential systematics. *Science 114:* 315, 1951.

Wright, S. Adaptation and selection. From *Genetics, Paleontology, and Evolution.* Ed.: G. Simpson, E. Mayr, and G. Stebbins. Princeton Univ. Press. pp. 365–389, 1949.

——. Personal communication. March 31, 1952.

THE CHAIN BLOCK DESIGN

W. J. Youden and W. S. Connor

National Bureau of Standards, Washington 25, D. C.

1. *Introduction.* The development of the design of experiments has been inspired largely by the needs of agriculture and biology. The first important steps were taken by R. A. Fisher and F. Yates, both of whom were on the staff of the Rothamsted Experiment Station. The methods devised by these workers are widely used in biology and agriculture and, to some extent, in other sciences.

We believe that one of the reasons for the delay in the adoption of experimental designs in the physical sciences is that the classical designs often do not meet the experimental situations encountered in these sciences. In fact, when one compares the experimental conditions which exist in the biological and agricultural sciences with the experimental conditions of the physical and chemical sciences, one is at once struck by certain fundamental differences. A difference which is of paramount importance is the magnitude of the errors of measurement in the two areas. In agricultural and biological experimentation the basic experimental material often is land or animals, and the variation over a field or between litters is likely to be large. To compensate for this large variation, the classical designs require repeated measurements to effect a reduction in the error of the estimates.

Physical measurements can be made with high precision and the experimental material usually is relatively homogeneous. It follows that it is not necessary to put great reliance on replication in order to achieve good results. Excellent estimates may be obtained from one measurement, or at most, two or three.

It is with the needs of the physical and chemical sciences in mind that we offer the Chain block design. This design is very flexible, and the number of replications is small. Thus, for a set of quantities to be compared, this new design calls for some of them to be measured once, and the others twice.

2. *Definition of the Chain Block Design.* Let us begin by considering a simple example of a Chain block design. Suppose 10 quantities are to be compared. Denote the quantities (traditionally, and hereafter, called treatments) by letters and arrange them in three blocks as follows.

BLOCK

1	2	3
A	C	E
B	D	F
C	E	A
D	F	B
a	b	c
d	e	

(2.1)

We observe that the treatments which are denoted by capital letters are replicated twice each, whereas the treatments which are denoted by small letters are replicated only once each. Further, A never occurs in a block without B, nor C without D, nor E without F. Thus, the treatments which are replicated twice fall into 3 groups, and these groups are the links in the chain of blocks. Treatments C and D link blocks 1 and 2, E and F link blocks 2 and 3, and A and B complete the chain by linking blocks 3 and 1.

In general, there are v treatments which are arranged in b blocks. A treatment is said to belong to class C_1 if it is replicated once, or to class C_2 if it is replicated twice. There are v_1 treatments in C_1 and v_2 treatments in C_2.

The treatments of C_1 are divided into b groups of n_{1j} treatments each, $(j = 1, \cdots, b)$, where a group is characterized by the property that the treatments of the j-th group occur in the j-th block, and nowhere else.

Similarly, the treatments of C_2 are divided into b groups of n_2 treatments each, $(n_2 \geq 1)$, where a group is characterized by the property that if a treatment of the group occurs in a block, then the other $n_2 - 1$ treatments of the group also occur in the block. Further, blocks j and $j + 1$ have in common the treatments of the j-th group, $(j = 1, \cdots, b; \text{mod } b)$, but no other treatments in common. No other two blocks have any treatments in common.

We shall denote the number of units in a block by k_i and let $\sum_{i=1}^{b} k_i = N$. From the above, the following relations hold:

$$
\begin{aligned}
v_1 + v_2 &= v, \\
v_1 + 2v_2 &= N, \\
\sum_{j=1}^{b} n_{1j} &= v_1 , \qquad (2.2) \\
bn_2 &= v_2 , \qquad \text{and} \\
n_{1j} + 2n_2 &= k_j .
\end{aligned}
$$

If we let G_{ij} , $(i = 1, 2; j = 1, \cdots, b)$, denote the treatments of the i-th class which are in the j-th group, then we may display the Chain block design as follows:

BLOCK

1	2	3	$\cdots$	b
G_{23}	G_{21}	G_{22}		$G_{2(b-1)}$
G_{21}	G_{22}	G_{23}		G_{2b}
G_{11}	G_{12}	G_{13}		G_{1b}

(2.3)

Since the average number of replicates is $(v_1 + 2v_2)/v$, we may refer to the Chain block design as a "$(v_1 + 2v_2)/v$ replicate" design. Of course,

$$1 < (v_1 + 2v_2)/v \leq 2. \qquad (2.4)$$

3. *Some Practical Aspects of the Design.* Perhaps the most striking aspect of the design is its flexibility. The numbers of treatments in C_1 and C_2 are largely at the disposal of the experimenter, as we shall see in the following example. Let $b = 4$ and $v = 22$, and let there be available only 10 plots per block. Then the following types of Chain block designs are possible.

DESIGN 1:

BLOCK

1	2	3	4
A	*C*	*E*	*G*
B	*D*	*F*	*H*
C	*E*	*G*	*A*
D	*F*	*H*	*B*
i	*m*	*q*	*t*
j	*n*	*r*	*u*
k	*o*	*s*	*v*
l	*p*		

(3.1)

In this design, $v_1 = 14$, $v_2 = 8$, $N = 30$, and there are 5 degrees of freedom for error. The number of degrees of freedom for error in the general case is given in Section 4.

DESIGN 2:

BLOCK

1	2	3	4
A	D	G	J
B	E	H	K
C	F	I	L
D	G	J	A
E	H	K	B
F	I	L	C
m	p	s	u
n	q	t	v
o	r		

(3.2)

In this design $v_1 = 10$, $v_2 = 12$, $N = 34$, and there are 9 degrees of freedom for error.

DESIGN 3:

BLOCK

1	2	3	4
A	E	I	M
B	F	J	N
C	G	K	O
D	H	L	P
E	I	M	A
F	J	N	B
G	K	O	C
H	L	P	D
q	s	u	v
r	t		

(3.3)

In this design $v_1 = 6$, $v_2 = 16$, $N = 38$, and there are 13 degrees of freedom for error. We have omitted one possibility, that for which $v_1 = 18$ and $v_2 = 4$, since in that design there is only one degree of freedom for error.

The choice of a design from among these three depends on several considerations. For one thing, it is clear that the estimates of the treatments of C_2 will have smaller variance than those of C_1. Hence, the

experimenter will have to decide whether there are approximately 8, 12, or 16 treatments which should be estimated with the smaller variance. Again, the experimenter must decide whether he needs 5, 9, or 13 degrees of freedom for error, and this decision will depend in part on the magnitude of the variance of a single determination or yield.

Although the definition of the Chain block design does not specify the distribution of the treatments of C_1 among the blocks, it seems desirable to distribute them as evenly as possible. For example, in Design 1 we have put either 3 or 4 treatments of C_1 in every block.

Now consider the difference between the estimates of any two treatments. The variance of this difference will depend on where the treatments are in the design. For example, consider the treatments of C_2 in Design 1. If the treatments are in the same group, as A and B, the variance of the difference between their estimates is the smallest variance in the design.

To explain this point further, let us think of the groups as arranged in a circle. Thus, around the circle we have groups 1, 2, $\cdots$, b, with b followed by 1. Now consider any two groups of C_2, say j and j'. Let the number of groups which lie between j and j', when counted in the shortest direction around the circle, be p. Then, the variance of the difference between the estimates of a treatment of j and one of j' varies directly with p. Thus, in Design 1, we have a larger variance for the difference between A and C than for A and B, and a still larger variance for the difference between A and E. Similar statements can be made for treatments in C_1, or for one treatment from C_1 and the other from C_2. The experimenter should, in so far as it is possible, be guided by these considerations when assigning treatments to groups.

4. *The Analysis in General.* The analysis which we shall give below can be rigorously justified, for example, by the theory of [1].

Let the typical yield be denoted by x_{ijuz}, where the first index refers to the class, the second to the group, the third to the treatment, and the fourth to the block. Thus, $i = 1$ or 2; $j = 1, \cdots, b$; $u = 1, \cdots, n_{ij}$; and $z = j$ or $j + 1$. Let t_{iju} denote the effect of the u-th treatment of the j-th group of C_i, and b_z denote the effect of the z-th block. Then we assume that

$$x_{ijuz} = t_{iju} + b_z + \epsilon_{iju}, \tag{4.1}$$

where ϵ_{iju} is a random variable with mean, zero, and with variance, σ^2. Further, we assume that ϵ_{iju} is independent of the corresponding random variable which is associated with any other yield.

Let the least squares estimate of a treatment or block effect be

denoted by putting a circumflex (^) over the corresponding parameter. Thus $\hat{t}_{iju}$ is the estimate of t_{iju} and $\hat{b}_z$ is the estimate of b_z .
Also, let

$$X_{2jz} = \sum_{u=1}^{n_2} x_{2juz} \qquad \text{and} \qquad D_j = X_{2ji} - X_{2j(j+1)} \ .$$

Then by imposing the restriction,

$$\sum_{z=1}^{b} b_z = 0, \tag{4.2}$$

we may estimate the effect of the j-th block by

$$(2n_2b)\hat{b}_i = \sum_{y=0}^{\alpha} (b - 2y - 1)(D_{i+y} - D_{i-y-1}), \tag{4.3}$$

where α is the largest integer which is less than or equal to $(b - 1)/2$. In (4.3) and in the formulas below the subscripts should be reduced, mod b.

The treatment estimates are easily found by using (4.3). Thus,

$$\hat{t}_{1ju} = x_{iju} - \hat{b}_i \tag{4.4}$$

and

$$2\hat{t}_{2ju} = x_{2juj} + x_{2ju(j+1)} - \hat{b}_j - \hat{b}_{j+1} \ . \tag{4.5}$$

To carry out the analysis of variance, we first compute the sum of squares due to error, which we shall denote by $S^2(e)$. For the j-th group of C_2 we form the differences,

$$d_{ju} = (x_{2juj} - x_{2ju(j+1)}), \qquad (u = 1, \cdots, n_2). \tag{4.6}$$

Now d_{ju} is an estimate of the difference between the j-th and $(j + 1)$st blocks. Next compute

$$S_j^2(e) = \frac{1}{2} \sum_{u=1}^{n_2} (d_{ju} - \bar{d}_{ju})^2 = \frac{1}{2}\left[\sum_{u=1}^{n_2} d_{ju}^2 - \frac{1}{n_2}\left(\sum_{u=1}^{n_2} d_{ju} \right)^2 \right], \tag{4.7}$$

where

$$\bar{d}_{ju} = \frac{1}{n_2} \sum_{u=1}^{n_2} d_{ju} \ .$$

The quantity $S_j^2(e)$ is an estimate of the sum of squares due to error, based on $(n_2 - 1)$ degrees of freedom. By computing a similar quantity for each group of C_2 , we obtain an estimate of the sum of squares due to error, based on $b(n_2 - 1)$ degrees of freedom.

The remaining degree of freedom is given by

$$Y = 1/(2bn_2)\left[\sum_{z=1}^{b} X_{2z(z+1)} - \sum_{z=1}^{b} X_{2zz}\right]^2. \tag{4.8}$$

Hence,

$$S^2(e) = \sum_{i=1}^{b} S_i^2(e) + Y. \tag{4.9}$$

We now compute the sum of squares due to blocks, uncorrected for treatments, and the total sum of squares; and then obtain the sum of squares due to treatments by subtraction. If we let B_i denote the sum of the yields in the j-th block and G denote the sum of all of the yields, then the sum of squares due to blocks uncorrected for treatments is

$$S^2(b) = \sum_{i=1}^{b} B_i^2/k_i - G^2/N, \tag{4.10}$$

and the total sum of squares is

$$S^2(T) = \sum x_{iiuz}^2 - G^2/N, \tag{4.11}$$

where the summation is over all values of the indices.

We obtain the sum of squares due to treatments by subtraction, thus:

$$S^2(t) = S^2(T) - S^2(e) - S^2(b). \tag{4.12}$$

The analysis of variance table is as follows:

ANALYSIS OF VARIANCE TABLE I

Due to	Degrees of Freedom	Sum of Squares	Mean Square
Treatments (corrected for blocks)	$v - 1$	$S^2(t)$	$s^2(t) = S^2(t)/(v - 1)$
Blocks (uncorrected)	$b - 1$	$S^2(b)$	$s^2(b) = S^2(b)/(b - 1)$
Error	$N - v - b + 1$	$S^2(e)$	$s^2(e) = S^2(e)/(N - v - b + 1)$
Total	$N - 1$	$S^2(T)$	

We may be interested in the sum of squares for the treatments of C_2 alone. If so, let

$$B_{2i} = X_{2(i-1)i} + X_{2ii} \quad \text{and} \quad G_2 = \sum_{i=1}^{b} B_{2i},$$

and compute the quantities,

$$S_2^2(b) = \frac{1}{2n_2}\left[\sum_{j=1}^{b} B_{2j}^2 - G_2^2/b\right] \tag{4.13}$$

and

$$S_2^2(T) = \sum x_{2iuz}^2 - G_2^2/2n_2b, \tag{4.14}$$

where the summation is over all values of the indices. Finally, the sum of squares due to treatments of C_2 , corrected for blocks, is

$$S_2^2(t) = S_2^2(T) - S^2(e) - S_2^2(b), \tag{4.15}$$

and the corresponding mean square is

$$s_2^2(t) = S_2^2(t)/(v_2 - 1). \tag{4.16}$$

We may want the sum of squares due to blocks corrected for treatments. This quantity can be found as follows. We compute the quantities

$$P_z = \tfrac{1}{2}(D_z - D_{z-1}), \tag{4.17}$$

($z = 1, \cdots, b$), and then find the sum of squares due to blocks corrected for treatments, $S^2(b)'$, by

$$S^2(b)' = \sum_{z=1}^{b} \hat{b}_z P_z . \tag{4.18}$$

The sum of squares due to treatments, uncorrected for blocks, $S^2(t)'$, is found by subtraction, thus,

$$S^2(t)' = S^2(T) - S^2(b)' - S^2(e). \tag{4.19}$$

We now write out the analysis of variance table.

ANALYSIS OF VARIANCE TABLE II

Due to	Degrees of Freedom	Sum of Squares	Mean Square
Treatments (uncorrected)	$v - 1$	$S^2(t)'$	$s^2(t)' = S^2(t)'/(v - 1)$
Blocks (corrected for treatments)	$b - 1$	$S^2(b)'$	$s^2(b)' = S^2(b)'/(b - 1)$
Error	$N - v - b + 1$	$S^2(e)$	$s^2(e)$
Total	$N - 1$	$S^2(T)$	

If we assume that the errors are normally distributed, we may carry out the usual F and t tests. It should be noticed that the ratio of $s_2^2(t)$ to $s^2(e)$ is F with $(v_2 - 1)$ and $(N - v - b + 1)$ degrees of freedom.

Let

$$f(z) = 1 + \frac{(2z - 1)b - 2z^2}{n_2 b}.$$

Then for any two treatments of C_2, say t_{2iu} and t_{2sw},

$$V(\hat{t}_{2iu} - \hat{t}_{2sw}) = \sigma^2, \quad j = s; \quad \text{or} \quad = [f(i)]\sigma^2, \quad j \neq s; \tag{4.20}$$

and for two treatments of C_1, say t_{1iu} and t_{1sw},

$$V(\hat{t}_{1iu} - \hat{t}_{1sw}) = [f(i) + (1 + 1/n_2)]\sigma^2, \tag{4.21}$$

where $i = j - s$, mod b. For a treatment of C_1 and a treatment of C_2, say t_{1iu} and t_{2sw},

$$V(\hat{t}_{1iu} - \hat{t}_{2sw}) = \frac{1}{2}\left(3 + \frac{b - 1}{n_2 b}\right)\sigma^2, \quad j = s, \quad \text{or}$$

$$= \left[f(i') + \frac{1}{2}\left(1 - \frac{b + 1 - 4i'}{n_2 b}\right)\right]\sigma^2, \quad j \neq s, \tag{4.22}$$

where $i' = j - s$, mod $(b + 1)$.

5. *An Example.* It is characteristic of many experimental situations in the physical sciences that the block is sharply defined. This contrasts with the arbitrary designation of a given land area as a block in agricultural field trials. For example, spectrographic determinations of the chemical elements may be carried out by the comparison of spectrum lines as recorded on a photographic plate. A limited number of exposures may be made on a plate which is then developed. Obviously, all determinations on one plate experience the same processing, and comparisons within a plate have been demonstrated to be more precise than comparisons between samples run on different plates. When the number of samples exceeds the capacity of a plate an opportunity is presented to use an arrangement to correct for block effects and to do this with a minimum amount of replication. The example chosen concerns a study of the nickel content of a large number of rods prepared from the same ingot. The study was made by B. F. Scribner of the Spectrochemistry Section of the National Bureau of

Standards. It is desired to detect possible differences in composition The rods were made from a selected portion of the ingot to insure a high degree of uniformity among them.

In the example, $b = 3$, $v_1 = 30$ and $v_2 = 12$. Hence, the design is a "$1\frac{2}{7}$ replicate".

We shall let iju, ($i = 1, 2$; $j = 1, 2, 3$; $u = 1, \cdots, n_{ij}$; $n_{ij} = 10$ or 4 according as $i = 1$ or 2) denote the u-th treatment of the j-th group of the i-th class. Then the treatments occur in the blocks as follows:

BLOCK

1	2	3	
231	211	221	
232	212	222	
233	213	223	
234	214	224	
211	221	231	
212	222	232	
213	223	233	(5.1)
214	224	234	
111	121	131	
.	.	.	
.	.	.	
.	.	.	
11(10)	12(10)	13(10)	

The amounts of nickel in the rods were recorded as logarithms to the base 10 of the ratio of the intensity of the nickel spectral line to the iron spectral line. For our purposes, these determinations have been coded by multiplying by 10^3, and then subtracting 170. We give below the coded amounts, and their corresponding symbols.

1 Symbol Amount	2 Symbol Amount	3 Symbol Amount
$x_{2311} = 8$	$x_{2112} = 4$	$x_{2213} = -1$
$x_{2321} = 7$	$x_{2122} = 3$	$x_{2223} = 0$
$x_{2331} = 14$	$x_{2132} = 10$	$x_{2233} = -3$
$x_{2341} = 9$	$x_{2142} = 6$	$x_{2243} = -8$
$x_{2111} = 13$	$x_{2212} = 5$	$x_{2313} = 1$
$x_{2121} = 15$	$x_{2222} = 7$	$x_{2323} = 5$
$x_{2131} = 12$	$x_{2232} = 2$	$x_{2333} = 2$
$x_{2141} = 9$	$x_{2242} = 6$	$x_{2343} = 0$
$x_{1111} = 11$	$x_{1212} = 10$	$x_{1313} = 5$
$x_{1121} = 5$	$x_{1222} = 9$	$x_{1323} = -1$
$x_{1131} = 17$	$x_{1232} = 6$	$x_{1333} = -3$
$x_{1141} = 14$	$x_{1242} = 7$	$x_{1343} = -6$
$x_{1151} = 12$	$x_{1252} = 6$	$x_{1353} = 2$
$x_{1161} = 13$	$x_{1262} = 4$	$x_{1363} = -2$
$x_{1171} = 14$	$x_{1272} = 7$	$x_{1373} = -2$
$x_{1181} = 12$	$x_{1282} = 7$	$x_{1383} = 0$
$x_{1191} = 8$	$x_{1292} = 9$	$x_{1393} = 1$
$x_{11(10)1} = 21$	$x_{12(10)2} = 10$	$x_{13(10)3} = 2$

(5.2)

We shall follow the method of analysis which was described above. Certain preliminary computations will be helpful. Thus, we obtain the following values for certain symbols which were defined in section 4:

$$X_{231} = 38 \quad X_{212} = 23 \quad X_{223} = -12$$
$$X_{211} = 49 \quad X_{222} = 20 \quad X_{233} = 8; \tag{5.3}$$

$$D_1 = 26, \quad D_2 = 32, \quad D_3 = -30; \tag{5.4}$$

$$\begin{aligned} d_{11} &= 9 & d_{21} &= 6 & d_{31} &= -7 \\ d_{12} &= 12 & d_{22} &= 7 & d_{32} &= -2 \\ d_{13} &= 2 & d_{23} &= 5 & d_{33} &= -12 \\ d_{14} &= 3 & d_{24} &= 14 & d_{34} &= -9 \end{aligned} \tag{5.5}$$

$$\sum_{u=1}^{4} d_{1u} = 26 \qquad \sum_{u=1}^{4} d_{2u} = 32 \qquad \sum_{u=1}^{4} d_{3u} = -30$$

$$\sum_{u=1}^{4} d_{1u}^2 = 238 \qquad \sum_{u=1}^{4} d_{2u}^2 = 306 \qquad \sum_{u=1}^{4} d_{3u}^2 = 278;$$

$$B_{21} = 87, \quad B_{22} = 43, \quad B_{23} = -4, \; G_2 = 126, \; \sum_{j=1}^{3} B_{2j}^2 = 9{,}434; \tag{5.6}$$

$$B_1 = 214, \; B_2 = 118, \; B_3 = -8, \; G = 324, \; \sum_{j=1}^{3} B_j^2 = 59{,}784; \tag{5.7}$$

and

$$\sum x_{2jus}^2 = 1{,}388, \qquad \sum x_{ijus}^2 = 3{,}862, \tag{5.8}$$

where the summations are over all values of the indices.

We estimate the effect of the first block by (4.3) and (5.4), thus,

$$24\hat{b}_1 = 2(56) = 112, \quad \text{or}$$

$$\hat{b}_1 = 14/3. \tag{5.9}$$

Similarly, $\hat{b}_2 = 1/2$ and $\hat{b}_3 = -31/6$. The sum of these estimates is zero, as it should be.

The estimate of the effect of t_{111} is found from (4.4) and (5.2) to be

$$\hat{t}_{111} = 11 - 14/3 = 19/3, \tag{5.10}$$

and the estimate of the effect of t_{211} is found from (4.5) and (5.2) to be

$$\hat{t}_{211} = \tfrac{1}{2}(13 + 4 - 14/3 - \tfrac{1}{2}) = 71/12. \tag{5.11}$$

Other treatment effects are estimated similarly. The sum of these estimates is 261, i.e., $G - \frac{1}{2}G_2$.

To estimate the sum of squares due to error we find from (4.7) and (5.5) that

$$S_1^2(e) = \tfrac{1}{2}[238 - \tfrac{1}{4}(26)^2] = 34.50, \tag{5.12}$$

$S_2^2(e) = 25.00$, and $S_3^2(e) = 26.50$; and from (4.8) and (5.3)

$$Y = \frac{1}{24}(77 - 49)^2 = 32.67, \tag{5.13}$$

so that by (4.9),

$$S^2(e) = 118.67. \tag{5.14}$$

From (4.10) and (5.7), the sum of squares due to blocks is

$$\begin{aligned} S^2(b) &= \frac{1}{18}(59{,}784) - \frac{1}{54}(324)^2 \\ &= 3{,}321.33 - 1{,}944.00 \\ &= 1377.33, \end{aligned} \tag{5.15}$$

and from (4.11), (5.7), and (5.8), the total sum of squares is

$$S^2(T) = 3{,}862 - \frac{1}{54}(324)^2 = 1{,}918.00. \tag{5.16}$$

Hence, by (4.12), (5.14), (5.15), and (5.16) the sum of squares due to treatments is

$$S^2(t) = 1{,}918.00 - 118.67 - 1{,}377.33 = 422.00. \tag{5.17}$$

The analysis of Variance Table I is as follows:

ANALYSIS OF VARIANCE TABLE I

Due to	Degrees of Freedom	Sum of Squares	Mean Square
Treatments (corrected for blocks)	41	422.00	10.29
Blocks (uncorrected)	2	1377.33	688.66
Error	10	118.67	11.87
Total	53	1918.00	

By (4.13) and (5.6),

$$\begin{aligned} S_2^2(b) &= \frac{1}{8}(9{,}434) - \frac{1}{24}(126)^2 \\ &= 1{,}179.25 - 661.50 \\ &= 517.75, \end{aligned} \tag{5.18}$$

and, by (4.14), (5.6), and (5.8),

$$\begin{aligned} S_2^2(T) &= 1{,}388.00 - \frac{1}{24}(126)^2 \\ &= 726.50. \end{aligned} \tag{5.19}$$

Hence, by (4.15), (5.14), (5.18), and (5.19), the sum of squares due to treatments of C_2 , corrected for blocks, is

$$S_2^2(t) = 726.50 - 118.67 - 517.75$$
$$= 90.08, \tag{5.20}$$

and by (4.16) the corresponding mean square is

$$s_2^2(t) = 90.08/11 = 8.19. \tag{5.15}$$

From (4.17), and (5.4),

$$P_1 = \tfrac{1}{2}(26 + 30) = 28.00, \tag{5.16}$$

$P_2 = 3.00$, and $P_3 = -31.00$. Thus by (4.18) and (5.9), the sum of squares due to blocks, corrected for treatments, is

$$S^2(b)' = 292.33, \tag{5.17}$$

and by (4.19), (5.14), (5.16), and (5.17), the sum of squares due to treatments, uncorrected for blocks, is

$$S^2(t)' = 1{,}918.00 - 118.67 - 292.33$$
$$= 1{,}507.00. \tag{5.18}$$

If the reader desires to do so, he may write out the Analysis of Variance Table II. The F ratio for blocks, corrected for treatments, is

$$F = 12.31, \tag{5.20}$$

which is significant at the .01 level of significance. Thus, we have been wise to remove the error due to blocks.

Since the treatment mean squares are not significant and in fact, are less than the error mean square, we might decide not to carry out t tests. However, for illustrative purposes, we shall compare the effects of 213 and 224. From (4.20), we find that the variance of the difference, $(t_{213} - t_{224})$, is $(13/12)\sigma^2$. Hence, using our estimate of σ^2, that is, $s^2(e)$, we obtain

$$t = \frac{|\, t_{213} - t_{224} \,|}{(13/12)^{1/2} s(e)} = \frac{|\, 3.17 - (-3.92) \,|}{3.731} = 1.90, \tag{5.21}$$

which is not significant.

REFERENCE

[1] Bose, R. C., "Least Squares Aspects of Analysis of Variance", Institute of Statistics of the University of North Carolina, Mimeograph Series 9, 1949.

DESIGN AND ANALYSIS OF TRIANGULAR SINGLY LINKED BLOCKS

K. R. Nair*

Institute of Statistics
The Consolidated University of North Carolina

Design.

Recently, the author (1950) called attention to the importance of enumerating partially balanced incomplete block (p.b.i.b.) designs involving as few as two replications of each treatment. In that and a subsequent note (Nair, 1951) several examples of p.b.i.b. designs involving only two replications and having either 2, 3 or 4 associate classes were given. Bose (1951) made an exhaustive study of two-replicate p.b.i.b. designs having 2 associate classes.

One of the two-replicate p.b.i.b. designs given in the author's (1950) paper consisted of a design having the parameters:

$$v = (1/2)\,p(p-1) \qquad k = (p-1) \qquad r = 2 \qquad b = p$$

$$\lambda_1 = 1 \qquad \lambda_2 = 0$$

$$n_1 = 2(p-2) \qquad n_2 = 1/2\,(p-2)(p-3)$$

$$p^1_{fg} = \begin{bmatrix} (p-2) & (p-3) \\ (p-3) & 1/2\,(p-3)(p-4) \end{bmatrix}$$

$$p^2_{fg} = \begin{bmatrix} 4 & 2(p-4) \\ 2(p-4) & 1/2\,(p-4)(p-5) \end{bmatrix}$$

This design is constructed by dualising the unreduced balanced incomplete block design in blocks of 2 plots for which the parameters v^*, k^*, r^*, b^* and λ^* are:

$$v^* = p, \quad k^* = 2, \quad r^* = (p-1), \quad b^* = (1/2)\,p(p-1), \quad \lambda^* = 1.$$

*Address: Forest Research Institute, Dehra Dun, India.

The dual design $v = b^*$, $k = r^*$, $r = k^*$, $b = v^*$ has the property that there is just one treatment common to every pair of blocks. It therefore belongs to the class of designs which Youden (1951) has recently used under the name *Singly Linked Blocks.*

Bose (1951) has shown that it is easy to write down the blocks of the design by using the scheme given below, illustrated for the case $p = 5$. The blocks are given by the rows (*or* columns).

*	1	2	3	4
1	*	5	6	7
2	5	*	8	9
3	6	8	*	10
4	7	9	10	*

SCHEME 1

Thus, the five blocks of the design are:

Block Number	Treatment Number			
(1)	1	2	3	4
(2)	1	5	6	7
(3)	2	5	8	9
(4)	3	6	8	10
(5)	4	7	9	10

It is interesting to note that the design can also be constructed from a simple geometrical configuration. This method consists in drawing p straight lines in such a way that they cut each other. There will be $(p - 1)$ points of intersection on each line and the total number of such points in the configuration will be $(1/2)p(p - 1)$. By considering the lines and points as analogous to blocks and treatments respectively, we obtain the design.

The following diagram illustrates the method for $p = 5$.

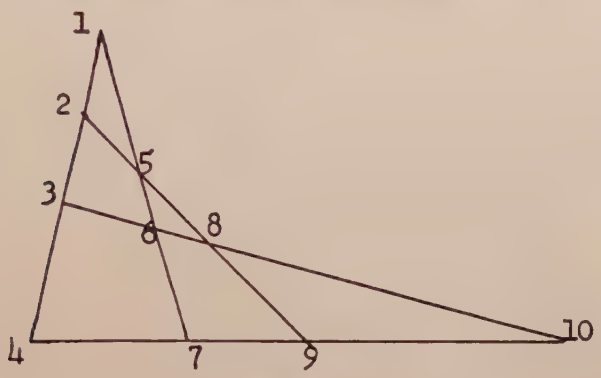

FIGURE 1.

When p is odd, the configuration can alternatively be presented artistically (like a star), as illustrated below for $p = 5$.

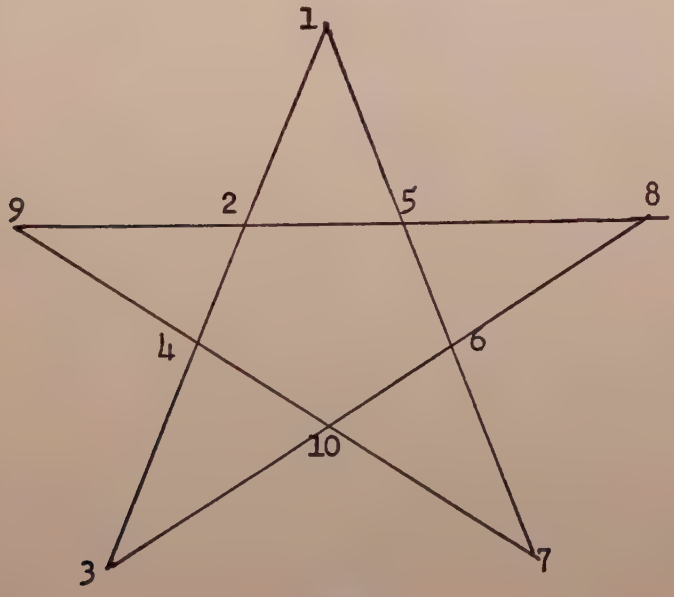

FIGURE 2.

One of the advantages of the scheme devised by Bose is that, in one stroke, it gives a simple method of constructing the design as well as a simple rule for picking out the first and second associates of each treatment. Thus, in Scheme 1, all treatments whose numbers appear in the

same row *and* column are first associates of each other. For instance, the 1st associates of treatment no. 1 are 2, 3, 4, 5, 6 and 7, while its second associates are 8, 9 and 10.

There is another way in which the association scheme of the design can be presented. Let us denote each treatment by two numbers (i, j) where i takes the values 1, 2, $\cdots$ $(p-1)$ and j takes the values $i+1, i+2, \cdots, p$, for a particular value of i. For instance the treatments 1, 2, $\cdots$ 10 of Scheme 1 may be written as follows:

(12)	(13)	(14)	(15)
	(23)	(24)	(25)
		(34)	(35)
			(45)

SCHEME 2

Treatment (ij) occurs in the ith and jth blocks of the design. Hence, treatments with one digit in common are 1st associates and treatments with no digit in common are 2nd associates.

Bose and Shimamoto (1952) have recently found that the *association scheme* of a number of p.b.i.b. designs having 2 associate classes, for which $v = 1/2\ p(p-1)$ but k not necessarily equal to $(p-1)$, can be represented in the same way as for the design: $v = 1/2\ p(p-1)$, $k = (p-1)$, illustrated in Scheme 1 for $p = 5$. They have given this type of p.b.i.b. designs the name: *Triangular Type.* In particular, the design for $v = 1/2\ p(p-1)$ and $k = (p-1)$ has been classified by them under the name *Triangular Singly Linked Blocks* (TSLB).

It is fairly obvious that Scheme 2 can be used to represent the association scheme of any triangular type design.

If the basic TSLB design involving only two replications of each treatment is repeated, say, l times in the experimental lay-out, so that there are l basic groups each having p blocks of $(p-1)$ plots, the resulting design may be considered as a partly resolvable p.b.i.b. design having the following parameters:

$$v = (1/2)\ p(p-1),\quad k = (p-1),\quad r = 2l,\quad b = lp,\quad \lambda_1 = l,\quad \lambda_2 = 0$$

and n_1 , n_2 , p^{s}_{fg} remaining the same as in the basic TSLB design.

The method of analysis for both the cases will be discussed in the

next two sections and a numerical example for the analysis of the basic design given in the last section of the paper.

Analysis of the Basic TSLB Design.

The statistical analysis of the data of an experiment laid out using triangular singly linked blocks can be performed by direct substitution in the general formulae for p.b.i.b. designs having 2 associate classes (See Nair, 1952). By using association Scheme 2, the resulting expressions for estimates of treatment effects, their variances and the components of the analysis of variance can be very much simplified.

The values of the quantitative character (x) under study for the two replications of treatment (ij) will, for the convenience of analysis, be denoted by x_{ij} for the plot belonging to the ith block and by x_{ji} for the plot belonging to the jth block. The whole data can then be presented as in Table 1 below.

Block Number (h)							Block Total ($x_{h.}$)
(1)	*	x_{12}	x_{13}	$\cdots$	$x_{1,p-1}$	x_{1p}	$x_{1.}$
(2)	x_{21}	*	x_{23}	$\cdots$	$x_{2,p-1}$	x_{2p}	$x_{2.}$
(3)	x_{31}	x_{32}	*	$\cdots$	$x_{3,p-1}$	x_{3p}	$x_{3.}$
.	.	.	.	$\cdots$	.	.	.
.	.	.	.		.	.	.
.	.	.	.		.	.	.
$(p-1)$	$x_{p-1,1}$	$x_{p-1,2}$	$x_{p-1,3}$	$\cdots$	*	$x_{p-1,p}$	$x_{(p-1).}$
(p)	x_{p1}	x_{p2}	x_{p3}	$\cdots$	$x_{p,p-1}$	*	$x_{p.}$
Total ($x_{.h}$)	$x_{.1}$	$x_{.2}$	$x_{.3}$	$\cdots$	$x_{.(p-1)}$	$x_{.p}$	$x_{..}$ = Grand Total

TABLE 1

Let T_{ij} denote the total value of x for treatment (ij). Then $T_{ij} = x_{ij} + x_{ji}$. Values of T_{ij} may be presented as in Table 2.

T_{12}	T_{13}	T_{14}	$\cdots$	T_{1p}
	T_{23}	T_{24}	$\cdots$	T_{2p}
			$\cdots$	$\vdots$
			$\cdots$	$\vdots$
				$T_{p-1,p}$

TABLE 2

The general formula (37) for estimating the effect of a treatment, with recovery of inter-block information, given in the author's (1952) paper simplifies for this particular design to the form

$$\bar{t}_{ij} = 1/2\left[T_{ij} - \mu\{(x_{i.} - x_{.i}) + (x_{j.} - x_{.j})\} - \frac{2x_{..}}{p(p-1)}\right] \tag{1}$$

where,

$$\mu = \frac{(w - w')}{pw + (p-2)w'} \tag{2}$$

The adjusted treatment *total* with recovery of inter-block information is therefore,

$$T_{ij} - \mu\{(x_{i.} - x_{.i}) + (x_{j.} - x_{.j})\} \tag{3}$$

Dividing by 2, we get the corresponding adjusted treatment *mean*.

To get the corresponding intra-block estimates, we have only to substitute $w' = 0$ or $\mu = 1/p$ in (1) and (3).

To estimate w and w' and from them the value of μ we have to perform the analysis of variance on the data of Table 1.

The total sum of squares, the treatment sum of squares (unadjusted for block effects) and the block sum of squares (unadjusted for treatment effects) are calculated in the usual way. They are:

$$\text{Total S.S.} = \sum_{j=i+1}^{p} \sum_{i=1}^{p-1} (x_{ij}^2 + x_{ji}^2) - \frac{x_{..}^2}{p(p-1)} \tag{4}$$

$$\text{Block S.S. (unadj.)} = \frac{1}{(p-1)} \sum_{h=1}^{p} x_{h.}^2 - \frac{x_{..}^2}{p(p-1)} \tag{5}$$

$$\text{Treatment S.S. (unadj.)} = (1/2) \sum_{j=i+1}^{p} \sum_{i=1}^{p-1} T_{ij}^2 - \frac{x_{..}^2}{p(p-1)} \tag{6}$$

It is easier in this design to first calculate the sum of squares for blocks (adjusted for treatment effects) than to calculate the sum of squares for treatments (adjusted for block effects). The former is given by the expression

$$\text{Block S.S. (adj.)} = \frac{1}{2p} \sum_{h=1}^{p} (x_{h\cdot} - x_{\cdot h})^2 \tag{7}$$

The latter is then calculated by subtracting (5) from the total of (6) and (7).

The analysis of variance is presented in Table 3.

TABLE 3. ANALYSIS OF VARIANCE

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Blocks (adj.)	$(p-1)$	See formula (7)	E_b
Treatments (unadj.)	$(1/2)(p+1)(p-2)$	See formula (6)	
Intra-block error	$(1/2)(p-1)(p-2)$	(4) − (6) − (7)	E_e
Total	(p^2-p-1)	See formula (4)	
Blocks (unadj.)	$(p-1)$	See formula (5)	
Treatments (adj.)	$(1/2)(p+1)(p-2)$	(6) + (7) − (5)	E_t

The variance ratio $F = E_t/E_e$ provides a test of significance of inter-block estimates of treatment effects.

The author (1944) had shown that for a non-resolvable incomplete block design (such as this) the estimates of w and w' are given by

$$w = \frac{1}{E_e}; \qquad w' = \frac{v(r-1)}{k(b-1)E_b - (v-k)E_e} \tag{8}$$

Substituting $v = (1/2)\ p(p-1)$, $k = (p-1)$, $r = 2$, $b = p$, we get

$$w' = \frac{p}{2(p-1)E_b - (p-2)E_e} \tag{9}$$

Hence, estimate of μ follows,

$$\mu = \frac{E_b - E_e}{pE_b} \tag{10}$$

For calculating the lowest significant differences among the values of $\bar{t}_{ij}$ given by (1) we have to calculate the variance of the difference between every pair of them. These pairs fall into two groups, namely,

those pairs which occurred together in the same block (i.e., those treatments with one digit common) and those pairs which did not occur together in the same block (i.e., those treatments with no digit common).

(i) Variance of the difference between two treatments with one digit common can be obtained by direct substitution in formula (39) of the author's (1952) paper. It simplifies to the form

$$\frac{1}{w}(1 + \mu) \tag{11}$$

(ii) Variance of the difference between two treatments with no digit common can be obtained by substitution in formula (40) of the author's paper. It simplifies to the form

$$\frac{1}{w}(1 + 2\mu) \tag{12}$$

(iii) The mean variance for differences for all pairs of treatments is derivable from formula (41) and simplifies to the form

$$\frac{1}{w}\left[1 + \frac{2(p-1)}{(p+1)}\mu\right] \tag{13}$$

By substituting $\mu = 1/p$ in (11), (12) and (13) we get the corresponding variances for intra-block estimates of treatment effects.

Analysis when the Basic TSLB Design is Repeated Several Times.

Let the basic TSLB design be repeated l times in l compact groups of p blocks each in the field, the p blocks of each group being independently randomized. In every basic group the blocks will be numbered in such a way that block no. (h) will be the one containing the $(p-1)$ treatments $(1, h)$, $(2, h)$, $\cdots$ $(h-1, h)$, $(h, h+1)$, $\cdots$ (h, p). These treatments will be randomized within this block.

Let x_{ijk} and x_{jik} be the values for treatment (ij) in the plots allotted to it in block numbers (i) and (j) of the kth basic group. Let

$$x_{ij.} = \sum_{k=1}^{l} x_{ijk} \; ; \qquad x_{ji.} = \sum_{k=1}^{l} x_{jik} \; .$$

We have to set up a table of values of $x_{ij.}$ and $x_{ji.}$ similar to Table 1 and calculate the marginal totals $x_{h..}$ and $x_{.h.}$; and the grand total $x_{...}$. The total for treatment (ij) will be $T_{ij} = x_{ij.} + x_{ji.}$.

The total for block no. (h) of the kth basic group will be $x_{h.k}$ and the total for the group will be $x_{..k}$.

Estimate of effect of treatment (ij) with recovery of inter-block

information is

$$\bar{t}_{ij} = \frac{1}{2l}\left[T_{ij} - \mu\{(x_{i..} - x_{.i.}) + (x_{j..} - x_{.j.})\} - \frac{2x...}{p(p-1)}\right] \quad (14)$$

where μ has the same value as in (2).

The corresponding adjusted treatment total is

$$T_{ij} - \mu\{(x_{i..} - x_{.i.}) + (x_{j..} - x_{.j.})\} \quad (15)$$

Dividing (15) by $2l$ we get the corresponding adjusted treatment mean.

To obtain intra-block estimates we have to replace μ by $1/p$.

In the analysis of variance the $(lp-1)$ degrees of freedom for blocks can be split up into three components of $(l-1)$, $(p-1)$ and $(l-1)(p-1)$ degrees of freedom. The first and third components are orthogonal to treatment comparisons. Their sums of squares are

$$\frac{1}{p(p-1)}\left(\sum_{k=1}^{l} x^2_{..k} - \frac{1}{l}x^2_{...}\right) \quad (16)$$

$$\frac{1}{(p-1)}\left(\sum_{1}^{p}\sum_{1}^{l} x^2_{h\cdot k} - \frac{1}{l}\sum_{1}^{p} x^2_{h..} - \frac{1}{p}\sum_{1}^{l} x^2_{..k} + \frac{1}{lp}x^2_{...}\right) \quad (17)$$

The expectation of (16) involves w, w' and the inter-group variance between blocks of size $p(p-1)$. It is therefore not available in estimating w'.

The expectation of (17) is $(l-1)(p-1)/w'$.

The second component is non-orthogonal with treatment comparisons. Its sum of squares after eliminating treatment effects is

$$\frac{1}{2lp}\sum_{1}^{p}(x_{h..} - x_{.h.})^2 \quad (18)$$

Using the results of Table VIIB of Bose and Shimamoto's paper we can show that the expectation of (18) is $1/2[(p)/w' + (p-2)/w]$.

To get the best estimate of w' we should combine (17) and (18). This has been done in the analysis of variance shown in Table 4.

Estimates of w, w' and μ in terms of E_b and E_e are given below

$$w = \frac{1}{E_e} \quad (19)$$

$$w' = \frac{2l(p-1) - (p-2)}{2l(p-1)E_b - (p-2)E_e} \quad (20)$$

$$\mu = \frac{l(E_b - E_e)}{lpE_b + (l-1)(p-2)E_e} \quad (21)$$

TABLE 4. ANALYSIS OF VARIANCE

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
(a) Basic groups	$l - 1$	$\frac{1}{p(p-1)}\left(\sum_1^l x_{\cdot\cdot k}^2 - \frac{1}{l}x_{\cdots}^2\right)$	
(b) Blocks within basic groups (adj.)	$l(p-1)$	$\frac{1}{2lp}\sum_1^p (x_{h\cdot\cdot} - x_{\cdot h\cdot})^2 + \frac{1}{(p-1)}\left(\sum_1^p\sum_1^l x_{h\cdot k}^2 - \frac{1}{l}\sum_1^p x_{h\cdot\cdot}^2 - \frac{1}{p}\sum_1^l x_{\cdot\cdot k}^2 + \frac{1}{lp}x_{\cdots}^2\right)$	E_b
(c) Treatments (unadj.)	$1/2(p+1)(p-2)$	$\frac{1}{2l}\sum_{j=i+1}^p\sum_{i=1}^{p-1} T_{ij}^2 - \frac{x_{\cdots}^2}{lp(p-1)}$	
(d) Intra-block error	$1/2(2lp-p-1)(p-2)$	$(e) - (a) - (b) - (c)$	E_e
(e) Total	$lp(p-1) - 1$	$\sum_{j=i+1}^p\sum_{i=1}^{p-1}\sum_{k=1}^l (x_{ijk}^2 + x_{jik}^2) - \frac{x_{\cdots}^2}{lp(p-1)}$	
(f) Blocks within basic groups (unadj.)	$l(p-1)$	$\frac{1}{(p-1)}\left(\sum_1^p\sum_1^l x_{h\cdot k}^2 - \frac{1}{p}\sum_1^l x_{\cdot\cdot k}^2\right)$	
(g) Treatments (adj.)	$1/2(p+1)(p-2)$	$(b) + (c) - (f)$	E_t

To calculate the variances of differences between adjusted treatment means with recovery of inter-block information we have to divide (11), (12) and (13) by l.

Numerical Example.

The data used in this example were made available to me by Dr. W. J. Youden of the Statistical Engineering Laboratory, National Bureau of Standards, Washington, D.C. from a Technical Report (unpublished) by John E. McKinney, George E. Decker and Frank L. Roth of the Bureau's Polymer Development Branch, Research and Development Division.

Physical properties of samples from 10 bales of a particular brand of synthetic rubber were measured in order to determine values of the properties for the lot and to measure the uniformity of the material throughout the lot.

Two compounded batches were prepared from samples of each of the 10 bales. Since there was not sufficient material from any one bale to make more than two tests and since it was not possible to test all the bales in a single day, the scheme in Table 5 was used.

TABLE 5. ALLOCATION OF BALES

Day	Bale Number			
(1)	1	2	3	4
(2)	1	5	6	7
(3)	2	5	8	9
(4)	3	6	8	10
(5)	4	7	9	10

It will at once be noticed that if 'days' and 'bales' denote 'blocks' and 'treatments' respectively, the scheme in Table 5 is identical to Scheme 1 for linked block designs for the case $p = 5$ discussed earlier.

Table 6 shows the values for one of the properties, namely, strain at 400 psi. Each value is the median for 3 specimens from a vulcanized sheet.

TABLE 6. DATA

Day	Strain (%)* at 400 psi			
(1)	35	20	13	25
(2)	16	16	21	27
(3)	10	5	20	15
(4)	26	24	37	31
(5)	21	16	20	17

*Coded—300

Using Scheme 2 for numbering the bales and Table 1 for presentation of the data we shall rearrange the data of Table 6 as shown in Table 7.

TABLE 7. RE-ARRANGED DATA WITH MARGINAL CALCULATIONS

Block No.						Total ($x_{h.}$)	($x_{.h}$)	$x_{h.} - x_{.h}$	$\mu(x_{h.} - x_{.h})$
(1)	*	x_{12} 35	x_{13} 20	x_{14} 13	x_{15} 25	93	73	+20	+2.94
(2)	x_{21} 16	*	x_{23} 16	x_{24} 21	x_{25} 27	80	80	0	0
(3)	x_{31} 10	x_{32} 5	*	x_{34} 20	x_{35} 15	50	93	−43	−6.32
(4)	x_{41} 26	x_{42} 24	x_{43} 37	*	x_{45} 31	118	71	+47	+6.91
(5)	x_{51} 21	x_{52} 16	x_{53} 20	x_{54} 17	*	74	98	− 24	−3.53
Total ($x_{.h}$)	73	80	93	71	98	415	415	0	0

The treatment totals are tabulated below:

(12)	(13)	(14)	(15)
51	30	39	46
	(23)	(24)	(25)
	21	45	43
		(34)	(35)
		57	35
			(45)
			48

(a) Total sum of squares $= (35)^2 + \cdots + (17)^2 - \frac{(415)^2}{20} = 1207.75$

(b) Block sum of squares (unadjusted) $= \frac{(93)^2 + \cdots + (74)^2}{4} - \frac{(415)^2}{20} = 626.00$

(c) Block sum of squares (adjusted) $= \frac{1}{10}[(+20)^2 + \cdots + (-24)^2] = 503.40$

(d) Treatment sum of squares (unadjusted) $= \frac{(51)^2 + \cdots + (48)^2}{2} - \frac{(415)^2}{20} = 504.25$

(e) Treatment sum of squares (adjusted) $= (c) + (d) - (b) = 381.65$

(f) Intra-block error sum of squares $= (a) - (c) - (d) \quad = (a) - (b) - (e)$

$= 200.10$

These sums of squares can now be presented in the following table of analysis of variance.

Source of variation	Degrees of Freedom	Sum of Squares	Mean Square
Blocks (adj.)	4	503.40	$125.85 = E_b$
Treatments (unadj.)	9	504.25	
Intra-block error	6	200.10	$33.35 = E_e$
Total	19	1207.75	
Blocks (unadj.)	4	626.00	
Treatments (adj.)	9	381.65	$42.41 = E_t$

The variance ratio $F = E_t/E_e = 1.27$ is not significant.
Estimate of μ is

$$\mu = \frac{E_b - E_e}{5E_b} = 0.1470$$

Each value of x in the hth block may now be adjusted by subtracting from it either $\mu(x_{h.} - x_{.h})$ or $1/p\ (x_{h.} - x_{.h})$ according as the adjustment desired is with or without recovery of inter-block information. The adjusted values for the former case are given below.

Block Number					
(1)	*	32.06	17.06	10.06	22.06
(2)	16.00	*	16.00	21.00	27.00
(3)	16.32	11.32	*	26.32	21.32
(4)	19.09	17.09	30.09	*	24.09
(5)	24.53	19.53	23.53	20.53	*

The corresponding adjusted treatment totals are:

(12)	(13)	(14)	(15)
48.06	33.38	29.15	46.59
	(23)	(24)	(25)
	27.32	38.09	46.53
		(34)	(35)
		56.41	44.85
			(45)
			44.62

The corresponding adjusted treatment means are:

(12)	(13)	(14)	(15)
24.0	16.7	14.6	23.3
	(23)	(24)	(25)
	13.7	19.0	23.3
		(34)	(35)
		28.2	22.4
			(45)
			22.3

In these triangular tables the easiest way to spot 1st associates is to remember that any two treatments with one digit in common would have occurred together in the same block e.g. treatments (12) and (25) occur together in block no. (2). Those with no digits common would not have occurred in the same block and are 2nd associates e.g. treatments (12) and (34).

(a) The variance of difference between means, adjusted with recovery of inter-block information, of two treatments occurring together in the same block (1st associates) is given by

$$\frac{1}{w}(1 + \mu) = 33.35 \times 1.147 = 38.25245$$

The lowest significant difference between adjusted means at 5 and 1% levels are

$$\text{L.S.D.}_{.05} = 2.447 \times \sqrt{38.25245} = 15.13$$

$$\text{L.S.D.}_{.01} = 3.707 \times \sqrt{38.25245} = 22.93$$

(b) The variance of difference between means, similarly adjusted, of two treatments not occurring together in the same block (2nd associates) is given by

$$\frac{1}{w}(1 + 2\mu) = 33.35 \times 1.294 = 43.15490$$

The lowest significant difference between adjusted means at 5 and 1% levels are

$$\text{L.S.D.}_{.05} = 2.447 \times \sqrt{43.15490} = 16.07$$

$$\text{L.S.D.}_{.01} = 3.707 \times \sqrt{43.15490} = 24.35$$

(c) The mean variance for differences for all pairs of adjusted treatment means is

$$\frac{1}{w}\left(1 + \frac{4}{3}\mu\right) = 33.35 \times 1.196 = 39.88660$$

If we are thinking of using one common L.S.D. for every pair of adjusted means, its value for 5 and 1% levels are

$$\text{L.S.D.}_{.05} = 2.447 \times \sqrt{39.88660} = 15.46$$

$$\text{L.S.D.}_{.01} = 3.707 \times \sqrt{39.88660} = 23.41$$

To get corresponding values of L.S.D. for intra-block adjusted treatment means we have only to replace μ by 1/5. These L.S.D. values

are strictly valid, but not the L.S.D. values calculated above for means adjusted with recovery of inter-block information owing to the fact that μ is estimated from the data and not known *a priori*.

Finally, I wish to express my thanks to Professor Gertrude M. Cox for her continued interest in this work which was done at the Institute of Statistics, The Consolidated University of North Carolina, Raleigh, during the tenure of a joint Fulbright and Smith-Mundt Fellowship awarded to the author by the U.S. Government. My thanks are also due to Dr. W. J. Youden for supplying me the experimental data for the numerical example.

REFERENCES

Bose, R. C. Partially balanced incomplete block designs with two associate classes involving only two replications. *Calcutta Statistical Association Bulletin, 3*, 120–125, 1951.

Bose, R. C. and Shimamoto, T. Classification and analysis of partially balanced incomplete block designs with two associate classes. *Journal of the American Statistical Association, 47*, 151–184, 1952.

Nair, K. R. The recovery of inter-block information in incomplete block designs. *Sankhya, 6*, 383–390, 1944.

Nair, K. R. Partially balanced incomplete block designs involving only two replications. *Calcutta Statistical Association Bulletin, 3*, 83–86, 1950.

Nair, K. R. Some two-replicate partially balanced designs. *Calcutta Statistical Association Bulletin, 3*, 174–176, 1951.

Nair, K. R. Analysis of partially balanced incomplete block designs illustrated on the simple square and rectangular lattices. *Biometrics, 8*, 122–155, 1952.

Youden, W. J. Linked blocks: A new class of incomplete block designs. *Biometrics, 7*, 124, 1951 (abstract).

SPLIT-PLOT HALF-PLAID SQUARES FOR IRRIGATION EXPERIMENTS[1]

WALTER C. JACOB[2]

The rapid developments in the field of experimental design have provided many designs which appear to be extremely complicated. It seems desirable to publish examples of the use of various designs so that other research men will have the benefit of knowing how the design actually worked in practice. It is the purpose of this paper to present an example of the use of a complex design in vegetable crops research. The reasons for the choice of the design and the various steps involved in the analysis and interpretation of the data are given in sufficient detail so that the non-statisticians will be able to utilize such a design.

One of the major projects of the Long Island Vegetable Research Farm is the determination of the fertilizer requirements of the various vegetable crops including potatoes. Recommendations have been made for fertilizing potatoes, but the advent of extensive irrigation of potatoes on Long Island raised the question as to how much difference there would be in fertilizer requirements of potatoes under irrigated conditions compared to potatoes grown without irrigation. Many new varieties of potatoes have been recently introduced and preliminary information indicated that some of the newer varieties had fertilizer requirements which differed considerably from the requirements of the standard varieties in use on Long Island.

The specific problem for which this experiment must furnish an answer could be stated as follows:

"What influence does irrigation have on the nitrogen, phosphorus, and potash requirements of different varieties of potatoes grown under Long Island conditions?" Thus there were five factors to be investigated, namely irrigation, nitrogen, phosphorus, potash and variety.

---

[1]Paper No. 310, Dept. of Vegetable Crops, Cornell University, Ithaca, New York.

[2]Professor of Vegetable Crops, Cornell University, Ithaca, New York.

SELECTION OF THE DESIGN

The area of land available for this experiment consisted of two large blocks each 214 feet by 260 feet in size. Irrigation was equally available over the whole area with the one restriction that all rows must parallel the irrigation lines and thus must run the 260 foot length of each block. Irrigation plots require large isolation borders and, since there was little interest in this experiment in testing the benefits of irrigation, it was decided to divide each block in half, one half to be irrigated and the other not. In this way any effect of irrigation would not be confounded with blocks and the estimation of fertilizer and variety interactions with irrigation could be viewed with more confidence. These whole plots consisted of 36 rows of potatoes each about 260 feet long. Four rows were left between plots for irrigation border areas.

Previous work (2)* has indicated the desirability of having at least one guard row on each side of a fertilizer plot to eliminate border effect. Thus two rows of each plot will have to be discarded and a minimum size plot would be three rows wide. To utilize the experimental area more efficiently and to provide some estimate of within plot variability, plots 4 rows in width were chosen as the smallest practical size. Using three levels each of nitrogen, phosphorus, and potash and three varieties, there were 81 treatments to be applied to each whole plot. Each whole plot had 9 sub-plots in the 36 row width and thus for 81 treatments each whole plot would have 9 plots down the length of the potato rows. Since each whole plot was over half an acre in size it was considered advisable to select an arrangement of the sub-plots so that soil variability within each whole plot could be removed at least partially from the estimate of error. Yates (5) has constructed a number of confounded arrangements in Latin squares. By confounding some of the interactions with the rows and columns of the squares, differences among the rows and columns can be eliminated from the experimental error. The 81 treatments can be arranged in a 9 x 9 quasi-Latin square by confounding portions of second and third order interactions. The potato rows corresponded in direction with the rows of the Latin square. Each plot was 28 feet long. One of the factors involved was variety and it was impractical to change variety at planting time for each plot. By applying only one variety to each row of the square we have the design described by Yates (5) as a half-plaid Latin square. Although this feature of the design was desirable from the standpoint of practicability in field operation it should be pointed out that considerable precision

*Numbers in parentheses refer to Literature Cited at end of paper.

was lost in the varietal comparisons. The comparison of varieties was not as important as some of the interactions between varieties and fertilizer factors so the loss in this comparison was felt to be justified by the increased precision available for the interactions and the saving in time and labor in installation of the experiment.

Figure 1 shows the construction form of the basic square before randomization (c. f. Yates (5)). This basic square was used in preparing the field layout for all four whole plots. Each plot was obtained from this basic square by a separate randomization of the rows and columns.

FIGURE 1. THE 9 × 9 HALF-PLAID LATIN SQUARE BEFORE RANDOMIZATION

		NPK "W" Confounded								
NPK "X" Confounded	*A*	111	213	312	122	221	323	133	232	331
	B	313	112	211	321	123	222	332	131	233
	C	212	311	113	223	322	121	231	333	132
	C	331	133	232	312	111	213	323	122	221
	A	233	332	131	211	313	112	222	321	123
	B	132	231	333	113	212	311	121	223	322
	B	221	323	122	232	331	133	213	312	111
	C	123	222	321	131	233	332	112	211	313
	A	322	121	223	333	132	231	311	113	212

Letters represent the three varieties.
Numbers represent levels of *N*, *P*, and *K* respectively.

Figure 2 shows the final field plan. The letters *A*, *B*, and *C* represent the three varieties which apply to the whole row of the square while the groups of three numbers indicate the levels of nitrogen, phosphorous, and potash, respectively. To confound interactions of factors having three levels each Yates (5) has prepared a formal subdivision of the degrees of freedom. The portion of the *N P K* interaction designated as *"W"* has been confounded with columns of the square and the *"X"* component has been confounded with rows. Thus those components of the *N P K V* and *N P K I* and *N P K V I* have also been confounded with rows or columns of the squares because of the split-plot half-plaid features. The actual planting was done with a special planter to be described in a later publication. The yields from each of the two middle rows for each of the 24 feet long plots were recorded separately. This gave information concerning the within plot variability or sampling error. This procedure would not always be necessary but sometimes knowledge of the sampling error is desirable for use in future work in planning experimental designs.

TABLE 1. YIELD OF U.S. No. 1 TUBERS FROM INDIVIDUAL PLOT ROWS. EXPRESSED IN POUNDS

	Block 1										Block 2									
										Total										Total
	26	28	34	31	33	33	28	24	34		38	26	30	28	28	30	29	24	22	
	28	29	41	40	46	40	36	29	40	600	34	20	23	24	24	27	27	28	20	482
	18	24	32	28	32	36	38	50	40		24	33	13	31	21	30	33	32	24	
	16	26	20	34	32	32	38	44	36	576	21	31	20	36	28	30	30	30	24	491
	36	41	38	32	25	30	28	33	34		24	30	18	37	24	32	26	25	26	
	28	29	40	26	30	27	28	34	36	575	22	25	20	30	28	30	26	26	24	473
	26	37	30	33	32	38	38	41	39		42	24	28	35	30	28	26	16	16	
	20	28	34	32	30	39	36	42	44	619	40	32	34	32	28	28	30	30	26	525
	19	23	37	42	36	40	44	38	48		28	18	16	21	33	28	28	26	24	
	24	20	44	38	36	36	48	48	46	667	36	23	28	27	33	30	30	32	26	487
	28	27	32	33	36	48	44	40	48		26	22	20	21	26	28	28	28	34	
	32	35	32	36	38	50	51	41	50	701	30	24	22	29	36	34	32	36	38	514
	27	40	30	27	34	36	38	41	34		20	32	12	35	37	42	38	35	32	
	30	38	36	29	28	35	29	31	27	590	20	29	12	34	33	34	34	33	37	549
	14	24	36	41	34	29	34	37	35		29	24	29	32	40	36	32	28	34	
	18	28	34	36	42	30	36	36	34	578	32	28	30	30	27	30	29	32	34	556
	16	30	35	32	35	40	28	27	36		34	24	30	40	40	34	34	35	36	
	17	39	42	35	42	47	30	38	24	593	36	38	24	38	28	30	31	36	34	602
Total	423	546	627	605	621	666	652	674	685	5499	536	483	409	560	544	561	543	532	511	4679

TABLE 1—*Concluded*

	Block 1										Block 2									
										Total										Total
	20	29	24	27	35	36	30	32	30		38	37	21	21	32	38	30	38	35	
	26	28	28	24	28	38	34	27	30	526	31	35	25	18	32	31	32	38	36	568
	31	32	33	30	30	26	28	36	34		36	36	38	44	26	36	40	44	42	
	30	32	35	28	28	29	27	35	30	554	33	34	42	32	18	29	46	46	47	669
	30	38	34	34	29	34	36	34	28		40	26	26	22	22	29	34	34	22	
	34	32	34	33	30	33	32	30	26	581	36	30	34	22	19	31	28	26	28	509
	28	28	32	30	25	34	27	24	28		32	43	46	40	22	28	49	46	48	
	29	34	30	26	24	32	32	30	34	527	38	46	26	30	36	35	47	44	46	702
	30	34	34	38	27	28	30	30	23		28	42	56	42	22	20	42	48	50	
	24	34	36	32	29	31	26	27	30	543	30	36	46	42	22	20	31	48	43	668
	17	22	29	24	21	26	26	24	22		35	46	54	41	28	28	44	46	57	
	24	28	25	22	30	32	30	32	28	462	33	46	31	30	37	38	42	42	52	730
	26	34	28	30	30	34	30	29	34		38	38	16	21	35	45	48	40	40	
	29	36	33	32	26	38	34	35	34	572	32	36	18	16	33	35	44	42	40	617
	35	32	32	28	28	30	28	30	34		40	27	32	36	43	47	40	53	34	
	35	32	32	28	28	30	28	30	34	554	40	26	28	31	45	41	46	50	34	693
	26	26	36	38	30	26	28	30	38		42	21	22	18	40	36	32	40	28	
	33	30	42	34	34	34	30	25	36	576	37	18	26	17	40	48	32	39	33	569
Total	507	561	577	538	512	571	536	540	553	4895	639	623	587	523	552	615	707	764	715	5725

FIGURE 2. RANDOMIZED PLAN USED IN FIELD PLANTING

Block 1
Irrigated

A	233	222	321	313	332	211	131	123	112
C	123	112	211	233	222	131	321	313	332
C	212	231	333	322	311	223	113	132	121
B	313	332	131	123	112	321	211	233	222
A	111	133	232	221	213	122	312	331	323
A	322	311	113	132	121	333	223	212	231
C	331	323	122	111	133	312	232	221	213
B	132	121	223	212	231	113	333	322	311
B	221	213	312	331	323	232	122	111	133

Isolation Aisle

C	112	222	131	313	233	321	332	123	211
A	222	332	211	123	313	131	112	233	321
C	231	311	223	132	322	113	121	212	333
B	213	323	232	111	331	122	133	221	312
C	323	133	312	221	111	232	213	331	122
B	121	231	113	322	212	333	311	132	223
A	133	213	122	331	221	312	323	111	232
A	311	121	333	212	132	223	231	322	113
B	332	112	321	233	123	211	222	313	131

Non-Irrigated

Isolation Aisle

Block 2
Non-Irrigated

B	222	131	233	123	211	321	313	332	112
C	121	333	132	322	113	223	212	231	311
A	112	321	123	313	131	211	233	222	332
A	323	232	331	221	312	122	111	133	213
C	213	122	221	111	232	312	331	323	133
A	231	113	212	132	223	333	322	311	121
B	133	312	111	331	122	232	221	213	323
B	311	223	322	212	333	113	132	121	231
C	332	211	313	233	321	131	123	112	222

Isolation Aisle

C	221	312	111	122	232	213	323	133	331
C	132	223	322	333	113	121	231	311	212
C	313	131	233	211	321	332	112	222	123
A	123	211	313	321	131	112	222	332	233
B	111	232	331	312	122	133	213	323	221
A	212	333	132	113	223	231	311	121	322
B	233	321	123	131	211	222	332	112	313
A	331	122	221	232	312	323	133	213	111
B	322	113	212	223	333	311	121	231	132

Irrigated

TABLE 2. YIELDS OF U. S. NO. 1 TUBERS TABULATED BY TREATMENTS

Irrigated

NPK	A			B			C		
	1	2	Total	1	2	Total	1	2	Total
111	43	68	111	65	58	123	56	46	102
112	74	63	137	62	82	144	50	62	112
113	64	71	135	59	39	98	56	44	100
121	74	88	162	52	64	116	70	65	135
122	76	53	129	58	44	102	66	39	105
123	53	70	123	65	34	99	34	50	84
131	64	58	122	64	37	101	68	56	124
132	69	85	154	32	61	93	67	69	136
133	43	86	129	60	40	100	62	76	138
211	73	89	162	74	68	142	52	44	96
212	81	68	149	77	48	125	64	89	153
213	72	103	175	69	73	142	61	69	130
221	80	60	140	33	93	126	72	69	141
222	57	96	153	83	80	163	64	60	124
223	95	65	160	70	35	105	57	70	127
231	98	66	164	76	79	155	70	86	156
232	81	67	148	87	78	165	67	64	131
233	54	94	148	83	70	153	62	60	122
311	62	86	148	69	84	153	55	90	145
312	92	88	180	77	84	161	71	72	143
313	71	72	143	46	80	126	94	76	170
321	75	70	145	77	74	151	76	41	117
322	60	109	169	73	79	152	58	80	138
323	94	88	182	77	96	173	78	62	140
331	86	80	166	67	102	169	57	71	128
332	79	90	169	65	92	157	76	60	136
333	98	92	190	70	80	150	78	76	154
	1968	2125	4093	1790	1854	3644	1741	1746	3487
									11,224

ANALYSIS OF DATA

Table 1 is a tabulation of the yield in pounds of U.S. No. 1 tubers from each of the two plot rows, arranged to correspond to the planting plan of Figure 2. This arrangement of the data is needed for the calculation of the total sum of squares and the sums of squares for rows and

TABLE 2—*Concluded*

Not Irrigated

NPK	*A*			*B*			*C*		
	1	2	Total	1	2	Total	1	2	Total
111	64	56	120	56	24	80	56	48	104
112	55	46	101	56	42	98	46	71	117
113	68	46	114	54	66	120	67	49	116
121	64	72	136	41	60	101	68	45	113
122	61	56	117	66	70	136	53	41	94
123	58	38	96	64	52	116	59	65	124
131	55	52	107	74	46	120	52	64	116
132	56	50	106	56	61	117	67	33	100
133	55	46	101	59	40	99	68	50	118
211	68	62	130	60	52	112	60	62	122
212	56	42	98	51	62	113	64	63	127
213	70	42	112	57	68	125	56	64	120
221	56	67	123	54	72	126	70	44	114
222	61	51	112	58	72	130	57	70	127
223	60	62	122	50	52	102	68	60	128
231	56	56	112	50	68	118	64	62	126
232	68	56	124	62	76	138	59	66	125
233	71	52	123	72	53	125	63	78	141
311	70	64	134	56	61	117	70	48	118
312	72	58	130	62	61	123	70	58	128
313	58	67	125	55	56	111	51	54	105
321	64	55	119	78	57	135	74	68	142
322	60	60	120	46	59	105	59	67	126
323	64	82	146	62	69	131	54	58	112
331	62	62	124	49	69	118	57	58	115
332	64	50	114	59	52	111	64	70	134
333	64	62	126	58	67	125	54	64	118
	1680	1512	3192	1565	1587	3152	1650	1580	3230
									9574

Block totals—10,394 and 10,404

columns in the analysis of variance, Table 4. The data in Table 2 are tabulated as in Table 1 by adding the two single row figures for each plot and identifying the plot treatments from Figure 2.

TABLE 3. TREATMENT TOTALS TABULATED FOR COMPLETE FACTORIAL ANALYSIS

NPK	Irrigated				Non-Irrigated				Total			Total
	A	*B*	*C*	Total	*A*	*B*	*C*	Total	*A*	*B*	*C*	
111	111	123	102	336	120	80	104	304	231	203	206	640
112	137	144	112	393	101	98	117	316	238	242	229	709
113	135	98	100	333	114	120	116	350	249	218	216	683
Total	383	365	314	1062	335	298	337	970	718	663	651	2032
121	162	116	135	413	136	101	113	350	298	217	248	763
122	129	102	105	336	117	136	94	347	246	238	199	683
123	123	99	84	306	96	116	124	336	219	215	208	642
Total	414	317	324	1055	349	353	331	1033	763	670	655	2088
131	122	101	124	347	107	120	116	343	229	221	240	690
132	154	93	136	383	106	117	100	323	260	210	236	706
133	129	100	138	367	101	99	118	318	230	199	256	685
Total	405	294	398	1097	314	336	334	984	719	630	732	2081
1T1	395	340	361	1096	363	301	333	997	758	641	694	2093
1T2	420	339	352	1112	324	351	311	986	744	690	664	2098
1T3	387	297	322	1006	311	335	358	1004	698	632	680	2010
Total	1202	976	1036	3214	998	987	1002	2987	2200	1963	2038	6201
211	162	142	96	400	130	112	122	364	292	254	218	764
212	149	125	153	427	98	113	127	338	247	238	280	765
213	175	142	130	447	112	125	120	357	287	267	250	804
Total	486	409	379	1274	340	350	369	1059	826	759	748	2333
221	140	126	141	407	123	126	114	363	263	252	255	770
222	153	163	124	440	112	130	127	369	265	293	251	809
223	160	105	127	392	133	102	128	352	282	207	255	744
Total	453	394	392	1239	357	358	369	1084	810	752	761	2323
231	164	155	156	475	112	118	126	356	276	273	282	831
232	148	165	131	444	124	138	125	387	272	303	256	831
233	148	153	122	423	123	125	141	389	271	278	263	812
Total	460	473	409	1342	359	381	392	1132	819	854	801	2474
2T1	466	423	393	1282	365	356	362	1083	831	779	755	2365
2T2	450	453	408	1311	334	381	379	1094	784	834	787	2405
2T3	483	400	379	1262	357	352	389	1098	840	752	768	2360
Total	1399	1276	1180	3855	1056	1089	1130	3275	2455	2365	2310	7130
311	148	153	145	446	134	117	118	369	282	270	263	815
312	180	161	143	484	130	123	128	381	310	284	271	865
313	143	126	170	439	125	111	105	341	268	237	275	780
Total	471	440	458	1369	389	351	351	1091	860	791	809	2460

TABLE 3. *Concluded*

NPK	Irrigated				Non-Irrigated				Total			Total
	A	*B*	*C*	Total	*A*	*B*	*C*	Total	*A*	*B*	*C*	
321	145	151	117	413	119	135	142	396	264	286	259	809
322	169	152	138	459	120	105	126	351	289	257	264	810
323	182	173	140	495	146	131	112	389	328	304	252	884
Total	496	476	395	1367	385	371	380	1136	881	847	775	2503
331	166	169	128	463	124	118	115	357	290	287	243	820
332	169	157	136	462	114	111	134	359	283	268	270	821
333	190	150	154	494	126	125	118	369	316	275	272	863
Total	525	476	418	1419	364	354	367	1085	889	830	785	2504
3T1	459	473	390	1322	377	370	375	1122	836	843	765	2444
3T2	518	470	417	1405	364	339	388	1091	882	809	805	2496
3T3	515	449	464	1428	397	367	335	1099	912	816	799	2527
Total	1492	1392	1271	4155	1138	1076	1098	3312	2630	2468	2369	7467
T11	421	418	343	1182	384	309	344	1037	805	727	687	2219
T12	466	430	408	1304	329	334	372	1035	795	764	780	2339
T13	453	366	400	1219	351	356	341	1048	804	722	741	2267
Total	1340	1214	1151	3705	1064	999	1057	3120	2404	2213	2208	6825
T21	447	393	393	1233	378	362	369	1109	825	755	762	2342
T22	451	417	367	1235	349	371	347	1067	800	788	714	2302
T23	465	377	351	1193	364	349	364	1077	829	726	715	2270
Total	1363	1187	1111	3661	1091	1082	1080	3253	2454	2269	2191	6914
T31	452	425	408	1285	343	356	357	1056	795	781	765	2341
T32	471	415	403	1289	344	366	359	1069	815	781	762	2358
T33	467	403	414	1284	350	349	377	1076	817	752	791	2360
Total	1390	1243	1225	3858	1037	1071	1093	3201	2427	2314	2318	7059
TT1	1320	1236	1144	3700	1105	1027	1070	3202	2425	2263	2214	6902
TT2	1388	1262	1178	3828	1022	1071	1078	3171	2410	2333	2256	6999
TT3	1385	1146	1165	3696	1065	1054	1082	3201	2450	2200	2247	6897
	4093	3644	3487	11224	3192	3152	3230	9574	7285	6796	6717	20798

The plot totals for the two replicates are then added and this total is transferred to a tabulation sheet set up as in Table 3. All of the subtotals necessary for the complete analysis of the factorial components are then obtained, by proper addition.

The partition of the degrees of freedom as given in Table 4 follows a combination of split-plot and half-plaid Latin square breakdowns. The whole plots are the four irrigation plots and the three degrees of freedom are divided one for irrigation, one for blocks and one for error

TABLE 4. ANALYSES OF VARIANCE

Source of Variance	Degrees of freedom	Sum of squares	Variance
Total	638	35,700.07	
Sub Total	323	31,002.07	
Irrigation (1)	1	4,201.40	4,201.40
Blocks	1	.16	
Error (a)	1	301.47	301.47
Rows	32	4,731.82	
Variety (V)	2	876.53	438.26
V I	2	983.38	491.69*
Error (b)	4	275.94	68.98
Columns	32	7,104.26	
Nitrogen (N)	2	3,980.51	1,990.26***
Phosphorus (P)	2	129.17	64.58
$N\ P$	4	112.28	28.07
Potash (K)	2	30.62	15.31
$N\ K$	4	103.00	25.75
$P\ K$	4	109.89	27.47
$N\ I$	2	884.60	442.30***
$P\ I$	2	152.02	76.01
$N\ P\ I$	4	15.70	3.92
$K\ I$	2	79.51	39.76
$N\ K\ I$	4	197.04	49.26
$P\ K\ I$	4	64.13	16.03
$N\ V$	4	162.14	40.54
$P\ V$	4	91.37	22.84
$N\ P\ V$	8	386.24	48.28
$K\ V$	4	117.25	29.31
$N\ K\ V$	8	361.44	45.18
$P\ K\ V$	8	190.74	23.84
$N\ I\ V$	4	223.84	55.96
$P\ I\ V$	4	63.53	15.88
$N\ P\ I\ V$	8	767.03	95.88**
$K\ I\ V$	4	201.99	50.50
$N\ K\ I\ V$	8	434.32	54.29
$P\ K\ I\ V$	8	277.88	34.74
Error (c)	148	5,526.72	37.34
Sampling Error	315	4,698.00	14.91

*Significant at $P \leq .05$
**Significant at $P \leq .01$
***Significant at $P \leq .001$

(a) (irrigation $\times$ blocks interaction). The degrees of freedom for the sub-plots are then partitioned as for four Latin squares. The four sets of 8 degrees of freedom for rows from each square are combined to give 32 degrees of freedom total for rows. Because of the half-plaid characteristic some of the row effects are confounded with variety effects. The 32 degrees of freedom for rows are divided into 2 for variety, 2 for $V\,I$, 4 for error (b) and the remainder. Error (b) is made up of 2 degrees of freedom for $V \times$ blocks and 2 for $V\,I \times$ blocks. There are also 32 degrees of freedom for columns, obtained in the same manner as for rows. The rest of the degrees of freedom are obtained in the usual manner of obtaining all the main effects and interactions. The only exception is $N\,P\,K$, $N\,P\,K\,V$, $N\,P\,K\,I$ and $N\,P\,K\,V\,I$. These are partially confounded with rows and columns and will not be calculated. Yates (5) gives a method for calculating the unconfounded portions of these interactions but the effort is rarely rewarded in cases of interactions of high order. Thus the confounded degrees of freedom of these $N\,P\,K$ interactions are included in rows and columns and the rest are lumped together with error (c). One other characteristic of this particular experiment was that the records from one row of potatoes were lost. In Table 1 these have been listed as duplicates of the other row of the plot in the next to the bottom row of plots of Block 1. These 9 degrees of freedom are thus lost from the sampling error and this accounts for 638 instead of 647 for total and 315 instead of 324 for sampling error.

The computations are done as follows:

1. The total sum of squares is obtained from the half-plot observations in Table 1 as follows:

$$\sum_{1}^{648} (26^2 + 28^2 + \cdots 33^2) - \frac{(20798)^2}{648}$$

2. The sub-total sum of squares is obtained from the plot totals in Table 2.

$$\sum_{1}^{324} \frac{(43^2 + 68^2 + \cdots 64^2)}{2} - \frac{(20798)^2}{648}$$

3. The sampling error is the difference between the sub-total and the total sums of squares.

4. The irrigation sum of squares is obtained from the totals in Table 3.

$$\frac{(11224)^2 + (9574)^2}{324} - \frac{(20798)^2}{648}$$

5. The block sum of squares is obtained from block totals in Table 2.

6. The four totals needed for the Block × Irrigation Interaction, which is error (a), are found in Table 1.

$$\frac{(5499)^2 + (4895)^2 + (4679)^2 + (5725)^2}{162} - \frac{(20798)^2}{648} - \text{S.S. for blocks and irrigation}$$

7. The row sum of squares is obtained from the row totals of Table 1.

$$\sum_{1}^{36} \frac{(600)^2 + (576)^2 + \cdots (569)^2}{18} - \frac{(20798)^2}{648} - \text{S.S. for blocks, irrigation and error } (a).$$

8. The column sum of squares is obtained from the column totals of Table 1 in the same manner.

9. Calculate the sum of squares for varieties using totals from Table 3.

10. The VI interaction is obtained by using the totals in Table 3.

11. Error (b) is calculated as follows, using totals from Table 2.

$$\sum_{1}^{12} \frac{(1968)^2 + (2125)^2 + \cdots (1580)^2}{54} - \frac{(20798)^2}{648} - \text{S.S. for } V, I, B, VI, \text{ error } (a).$$

12. The sums of squares for N, P, K and the interactions given in Table 4 are calculated according to standard factorial procedures (5) using appropriate totals from Table 3. For example:

$$\text{S.S. for } N = \frac{(6201)^2 + (7130)^2 + (7467)^2}{216} - \frac{(20798)^2}{648}$$

$$\text{S.S. for } NP = \sum_{1}^{9} \frac{(2032)^2 + (2088)^2 + \cdots (2504)^2}{72} - \frac{(20798)^2}{648} - \text{S.S. for N and P}$$

$$\text{S.S. for } NPI = \sum_{1}^{18} \frac{(1062)^2 + (970)^2 + \cdots (1085)^2}{36} - \frac{(20798)^2}{648} - \text{S.S. for } N, P, NP, I, NI, PI$$

13. Experimental error (c) is obtained by subtraction of the components already computed from the sub-total sums of squares.

Further details of computation may be found in Yates (5), Snedecor (4) under general description of analysis in factorial experiments, and

Cochran and Cox (1) Chapter 8. Significance tests and levels of significance can be found in Snedecor (4).

The analysis may be carried further to single degree of freedom comparisons as described by Yates (5) and Snedecor (4). This step is used in interpretation but will not be described here.

INTERPRETATION OF RESULTS

The significance of the various main effects and interactions must be determined by comparison of the variances with the proper error variance. Irrigation can be tested only with error (a) and no significance is indicated. Since the design selected had a low precision for testing the effect of irrigation, small differences could not be detected in this experiment. The large variance due to irrigation might indicate that some advantage could be gained by analyzing the irrigated and non-irrigated sections separately. Since the separate analyses did not in this case change the interpretations they are omitted for simplicity.

Even with the few degrees of freedom available the variety irrigation interaction was significant. The interpretation of this interaction is facilitated by plotting the means as in Figure 3. From this it is seen that without irrigation all varieties yielded about the same but with irrigation Variety A was much higher than B or C and there was no difference between B and C. Thus the interaction was the marked response of Variety A to irrigation compared to the lesser response of B and C.

There was a significant response to nitrogen. Examination of the means (Table 5) reveals that there was a significant increase with each

TABLE 5. RESPONSE TO NITROGEN APPLICATIONS

Level of Nitrogen	Mean yield in lbs. per plot row
N_1	28.71
N_2	33.01
N_3	34.57

increment of N even though the response to the second increment was much less than the response to the first increment. A more convenient method of analyzing the N effect is to break out the two individual degrees of freedom. This is logically done by the linear and quadratic terms as described by Yates (5). This divides the 3980.51 sum of squares for N into 3710.08 for linear and 270.41 for quadratic. Both

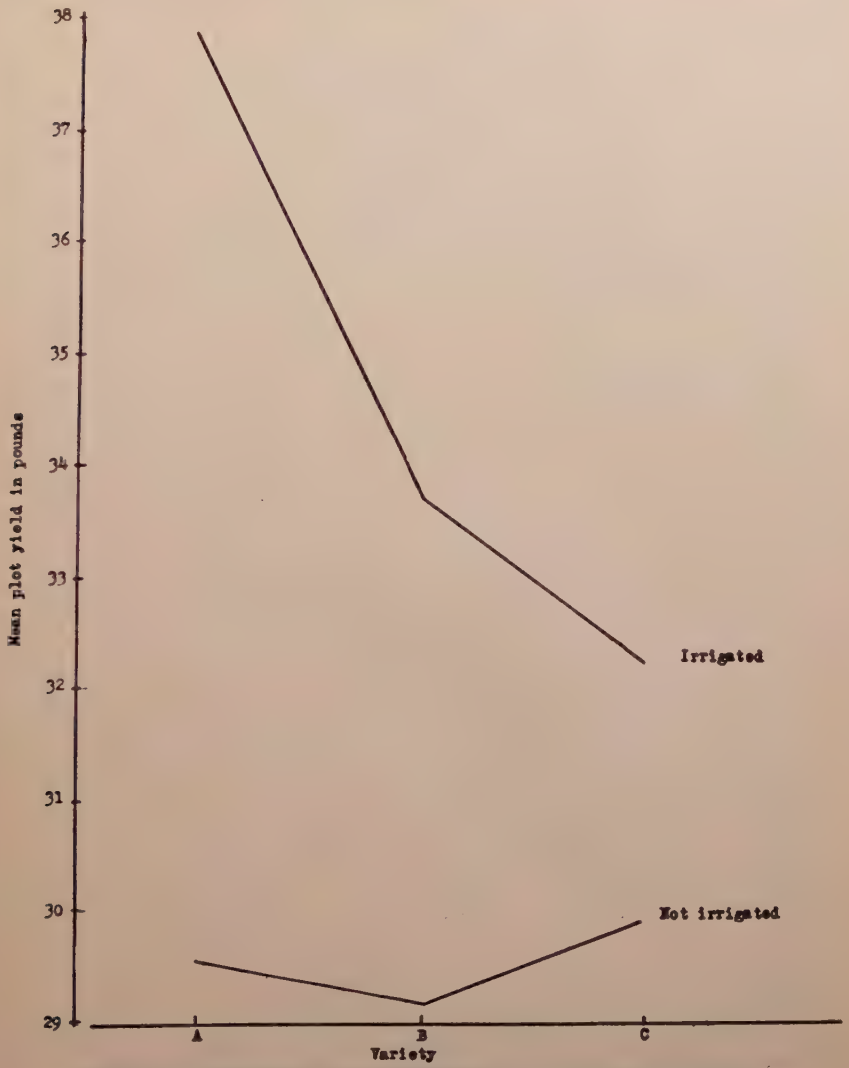

FIGURE 3. INFLUENCE OF IRRIGATION ON YIELD RELATIONSHIPS OF THE VARIETIES

terms are significant when compared by F test to the error (c). This means that although there was a marked linear response to N the second increment was less effective than the first.

Further examination of the analysis of variance indicates a significant $N\,I$ interaction. Comparison* of this variance with the variance

*This comparison is valid on the assumption that the levels of N and I are samples of an infinite population of levels. If these are considered as fixed effects such a comparison is not valid.

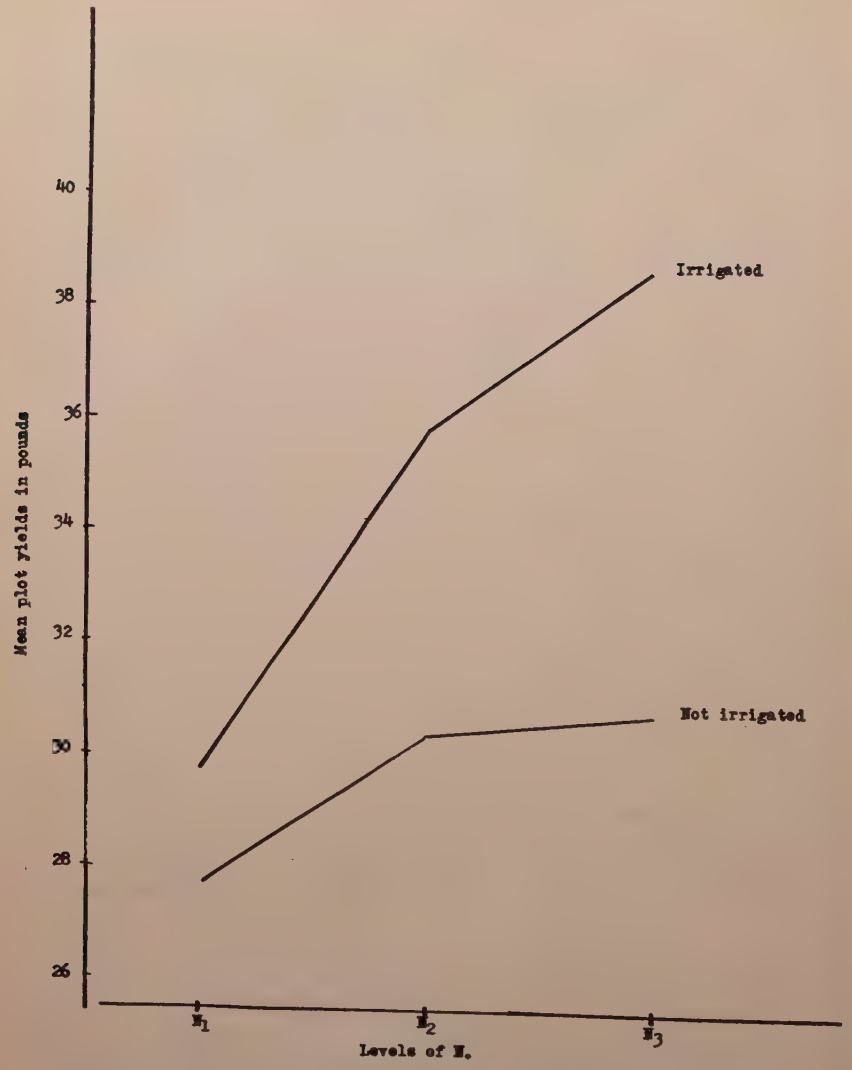

FIGURE 4. INFLUENCE OF IRRIGATION ON RESPONSE TO NITROGEN

for N shows that N variance is not significantly larger than the $N\,I$ interaction variance. Thus the effect of N must be interpreted only in terms of the irrigation level. The means for the $N\,I$ interaction are plotted in Figure 4. Here it is seen that the response to N was much greater with irrigation than without. There seems to be a similar tendency toward less response from the second increment at both irrigation levels. If the interaction is broken down to one degree of freedom

for linear N by irrigation and one for quadratic N by irrigation it will be found that the variance for linear $N \times$ Irrigation is 878.35 and for quadratic $N \times I$ is 6.25. Obviously the latter is non-significant indicating the same reduction in response to the second increment of N with and without irrigation. The linear responses indicate a much greater response to N with irrigation than without. Even though the reductions in response from the second increment of N were the same with and without irrigation it is evident that with irrigation the limit of economic N application will be much greater than without because of the greater response in each increment.

The significance of the $N\ P\ I\ V$ interaction poses some new problems. First the $N\ I$ variance and the $V\ I$ variance should be compared with $N\ P\ I\ V$ variance to determine whether $N\ I$ and $V\ I$ can still be interpreted as above in spite of the influence of P and V in them. $N\ I$ variance is significantly larger than $N\ P\ I\ V$ variance so the above interpretation can stand. $V\ I$ variance is not different from $N\ P\ I\ V$ so the $V\ I$ interpretation must be altered. The $N\ P\ I\ V$ means are plotted in Figure 5 for ease of interpretation. Without irrigation there was little response to N. However, with irrigation there was response to N and the degree of response varied with variety and phosphorus level. Variety A responded best to N at P_3, Variety B responded best at P_2 or P_3, the response at P_2 being linear, but at P_3 a reduced effect was found for the second increment of N. Variety C responded best to N at P_1. Thus the response of the varieties to N with irrigation must be interpreted with respect to specific varieties at definite levels of P.

One other point of interest is the comparison of the sampling error and the experimental error (c). Using the notation found in Snedecor (4) it is found that $S_E^2 = 11.21$ and $S_S^2 = 14.91$. The estimate of the variance of a treatment mean is:

$$S_T^2 = \frac{S_E^2}{R} + \frac{S_S^2}{Rk} = \frac{11.21}{R} + \frac{14.91}{Rk}$$

where R is the number of plots of the treatment and k is the number of rows per plot. It is evident that increasing replications or size of plot will both reduce the treatment mean variance. However, it would seem that increasing replications will be more advantageous in this case since R is present in both terms of the equation and the two terms are the same size. If S_S^2 was markedly greater than S_E^2 then there also would be some advantage in increasing the number of rows per plot. Whether additional precision is desired depends on the objectives of the experiment. For this experiment, destined to be continued for several years, the degree of precision seems adequate.

DISCUSSION AND SUMMARY

That this design has succeeded in removing a large portion of field variability was confirmed by the very large components of variance associated with columns and with rows exclusive of varieties. The gain in precision may be calculated. Assuming that the irrigation plots would have to be split for varieties and the fertilizer treatments then randomized on each variety plot, and assuming that the N P K inter-

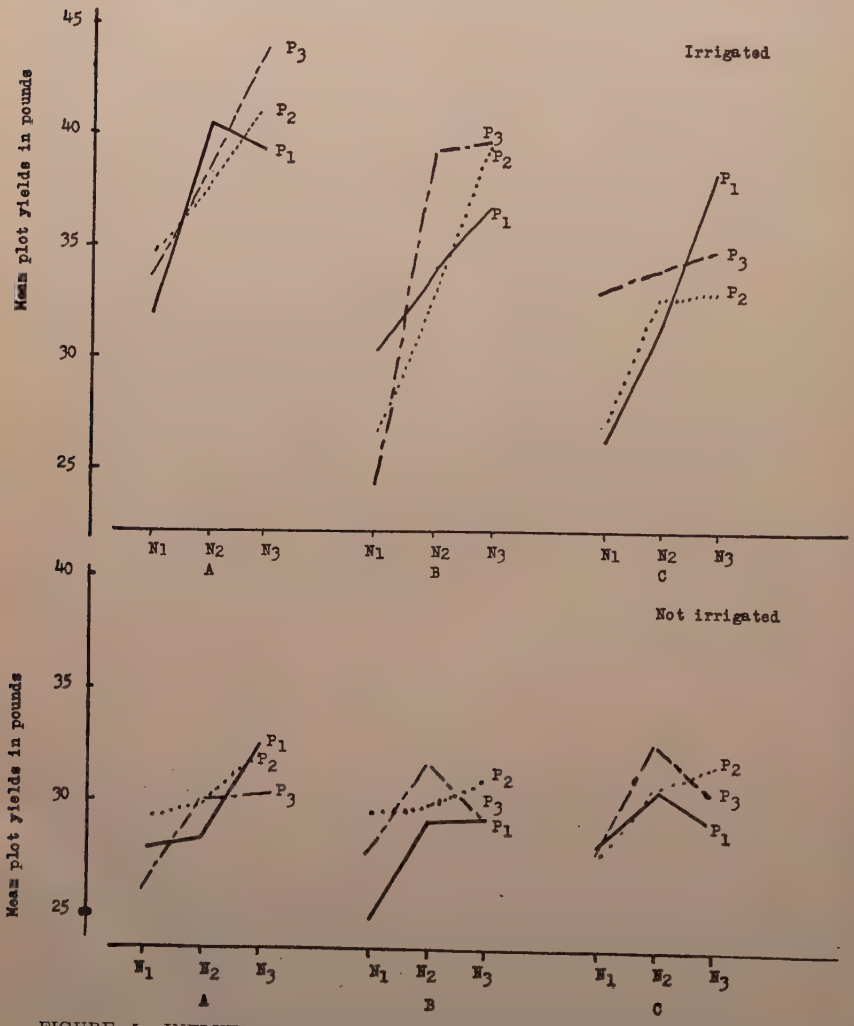

FIGURE 5. INFLUENCE OF IRRIGATION ON THE EFFECT OF PHOSPHORUS ON THE RESPONSE OF VARIETIES TO NITROGEN

actions are not significant, the error variance for such a design would be the rest of the rows plus columns plus error (c). This gives a new error variance of approximately 75. Dividing 37 by 75 we obtain 49 per cent as the efficiency of an unconfounded arrangement. The reciprocal of this is 204 per cent and the gain in information due to confounding is 104 per cent.

This paper is a description of the split-plot half-plaid Latin square in a field irrigation experiment with fertilizer and potato varieties. Some of the major considerations which led to the selection of the design are discussed. A description is given of the method of analysis of the data and some of the interactions are discussed. The use of the half-plaid Latin square feature has resulted in a gain of 100 per cent in information compared to the split-plot design. For a complete interpretation of three years' data from this experiment the reader is referred to (3).

LITERATURE CITED

1. Cochran, W. G. and G. M. Cox. *Experimental Designs*, 1950.
2. Jacob, W. C. Importance of Border Effect in Certain Kinds of Field Experiments with Potatoes. *Amer. Soc. for Hort. Sci. Proc. 37:* 866–870, 1939.
3. Jacob, W. C., et al. The Influence of Irrigation on the Nitrogen, Phosphorus and Potash Requirements of Different Potato Varieties. *Amer. Pot. Jour. 26:* 241–255, 1949.
4. Snedecor, G. W. *Statistical Methods.* Iowa State College Press, 1946. Fourth Edition.
5. Yates, F. The Design and Analysis of Factorial Experiments. *Imp. Bur. of Soil Sci. Tech. Com. No. 35*, 1937.

FITTING THE NEGATIVE BINOMIAL DISTRIBUTION TO BIOLOGICAL DATA

C. I. Bliss

The Connecticut Agricultural Experiment Station and Yale University

Note on the Efficient Fitting of the Negative Binomial

R. A. Fisher

Department of Genetics Cambridge

In studying the occurrence of plants and animals in nature, the number of individuals may be counted in each of many equal units of space or time. The original counts can be summarized in a frequency distribution, showing the number of units containing $x = 0, 1, 2, 3, \cdots$ individuals of a given species. If every unit in the series were exposed equally to the chance of containing the organism, the distribution would follow the Poisson series, each unit having the population mean as its expected frequency. It is easy to test whether the variation in the number of individuals per unit agrees with this hypothesis. Since the expected variance of a Poisson distribution is equal to its mean, the observed variance s^2, multiplied by the degrees of freedom n, may be divided by the sample mean $\bar{x}$ to obtain $\chi^2 = ns^2/\bar{x}$. More often than not χ^2 is significantly larger than its expectation, not only in distributions of plants and animals in nature but even in the laboratory.

A number of distributions have been devised for series in which the variance is significantly larger than the mean (2, 11, 21), frequently on the basis of more or less complex biological models. In the present paper this characteristic will be called "over dispersion". Perhaps the first of these was the negative binomial, which arose in deriving the Poisson series from the point binomial (27, 32) although it had been formulated in 1714 (2). Comparisons of expected and observed distributions have shown its wide applicability to biological data. The relative ease with which the negative binomial can be computed and

other desirable properties that have been described by Fisher (12) and by Anscombe (2) have now been supplemented by a practicable maximum likelihood estimation of its parameters described in the note by Sir Ronald Fisher at the end of the present paper. Here we will consider some characteristics of the negative binomial, estimates of its parameters and their precision, the calculation of the expected frequencies for a given sample, and its applicability to biological data.

The Negative Binomial Distribution. The negative binomial distribution is completely defined by two parameters, the arithmetic mean m and a positive exponent k. It is so called by analogy with the positive binomial distribution, $(q + p)^{n'}$, where n' is the number of individuals in a group and q and p are the expected proportions in two contrasting categories with $q + p = 1$. In computing the statistics of the positive binomial from distributions of yeast cells counted with a haemocytometer, Student (27) observed that two of his series gave negative values for p and n' but nevertheless fitted his observations very well. These and other cases (32) were described by the negative binomial, $(q - p)^{-k}$, where $p = m/k$ and $q = 1 + p$. By expansion of this expression, the probability P_x that an observational unit x will contain 0, 1, 2, $\cdots$ individuals is

$$P_x = \frac{(k + x - 1)!}{x!(k - 1)!} \cdot \frac{R^x}{q^k}, \tag{1}$$

where $R = p/q = m/(k + m)$. The probability for a given x is multiplied by N, the total number of units counted, to obtain the expected frequency (ϕ) of units with x individuals. The curve defined by the P_x's (or ϕ's) is unimodal, so that in fitting the negative binomial to an observed distribution any apparent bimodality (or multimodality) is attributed to random sampling.

The negative binomial is an extension of the Poisson series in which the population mean m, the parameter of the Poisson distribution, is not constant but varies continuously in a distribution proportional to that of χ^2. As the variance of a negative binomial approaches the mean, or the over-dispersion decreases, $k \to \infty$ and $p \to 0$. Under these conditions it can be shown (13) that the distribution converges to that for the Poisson, $P_x = e^{-m}(m^x/x!)$. Conversely, if the over-dispersion increases sufficiently, $k \to 0$. If we disregard the number of units containing no individuals, the negative binomial then converges to Fisher's logarithmic series (13), which describes effectively the apparent abundance of different species. Thus the limiting values of the exponent lead to distributions of importance in biology.

An example of the negative binomial is provided by counts of the number of European red mites on apple leaves, for which I am indebted to Dr. Garman of The Connecticut Agricultural Experiment Station (15). On July 18, 1951, 25 leaves were selected at random from each of six McIntosh trees in a single orchard receiving the same spray treatment, and the number of adult females counted on each leaf. The frequency distribution of mites on the 150 leaves is given in the first two columns of Table 1.

TABLE 1

Fitting the negative binomial to counts of red mites on apple leaves, data of P. Garman (15).

No. of mites per leaf x	No. of leaves observed f	Accumulated frequencies A_x	Expected frequencies ϕ	$\frac{(f-\phi)^2}{\phi}$
0	70	80	69.49	.004
1	38	42	37.60	.004
2	17	25	20.10	.478
3	10	15	10.70	.046
4	9	6	5.69	1.925
5	3	3	3.02	
6	2	1	1.60	.027
7	1		.85	
8+	0		.95	
Total	150 = N		150.00	2.484 = χ^2

$S(fx) = 172, \quad S(fx^2) = 536, \quad S(fx^3) = 2170$
$\bar{x} = 1.14667$ (Eq. 2), $\quad s^2 = 2.27365$ (Eq. 4)

As a test of agreement with the Poisson series, the observed variance (s^2) has been computed from the basic sums beneath the table. It was nearly twice as large as the mean. From $\chi^2 = 149 \times 2.27365/1.14667 = 295.44$ with 149 degrees of freedom, P is less than .001. The over-dispersion was clearly far too large for the Poisson series. By contrast, the frequencies expected by the negative binomial, shown in the fourth column of Table 1, reproduced the observed values very closely. In the next sections, the detailed computation of the negative binomial will be illustrated with this example.

The Statistics of the Negative Binomial. The parameters m and k of the negative binomial are estimated from the frequency distribution of

a sample by the statistics $\bar{x}$ and $\hat{k}$. The mean is estimated efficiently from the frequency f of units at each x as

$$\bar{x} = S(fx)/N \tag{2}$$

For the distribution in Table 1 its numerical value was $\bar{x} = 1.146667$.

The exponent k is more difficult. Two approximations (methods 1 and 2) are available, each with a relatively high efficiency for suitable combinations of m and k (1, 2). With the maximum likelihood solution (method 3) in the note appended to this paper, either estimate may be used as a first step toward a fully efficient fitting.

(1) The original and simplest solution for the statistic k is that based upon the first and second moments (12, 32). It is determined from the mean and variance s^2 of the sample as

$$\hat{k}_1 = \frac{\bar{x}^2}{s^2 - \bar{x}}, \tag{3}$$

where, as usual, the variance is

$$s^2 = \frac{S(fx^2) - S^2(fx)/N}{N - 1} \tag{4}$$

The efficiency of the moment solution as defined by Fisher (12) has been plotted by Anscombe (1, 2) for different combinations of m and k. It has an efficiency of 90 percent or more for small values of m when $k/m > 6$, for large values of m when $k > 13$, and for m in the intermediate zone when $(k + m)\ (k + 2)/m \geq 15$. In practice the population values m and k are replaced necessarily by the statistics $\bar{x}$ and $\hat{k}$.

In our example, the estimation of $\hat{k}_1$ leads by Eq. 3 to $\hat{k}_1 = (1.14667)^2/(2.27365 - 1.14667) = 1.16670$. If $\hat{k}_1$ has been estimated with an efficiency of 90 percent or better, the inequality $2.313 \times 3.167/1.147 \geq 15$ should hold, but since $6.39 \not> 15$, this method is not suitable for the present series.

(2) An alternative estimate (1, 2) is based upon the ratio of the total number of units in the sample (N) to the number of units without organisms (f_0). From Eq. 1 the expected probability for $x = 0$ is $P_0 = 1/q^k$, and if P_0 is replaced by the proportion observed in the zero class, we have $f_0/N = 1/q^k$. Since $q = 1 + m/k$, an iterative solution is necessary. The required estimate of $\hat{k}_2$ is that which balances the equation

$$\hat{k}_2 \log (1 + \bar{x}/\hat{k}_2) = \log (N/f_0) \tag{5}$$

The left side of Eq. 5 is computed twice, with different trial values k', one giving a larger and the other a smaller product than the con-

stant term on the right of the equality. Interpolating between these two products for $\log(N/f_0)$ leads to a first approximation of $\hat{k}_2$, with which the process can be continued until $\hat{k}_2$ has the required precision. To estimate k with an efficiency of 90 percent or more, at least 1/3 of the units must be empty, but if the mean is less than 10, enough more empty units are required to satisfy the inequality $(m + 0.17)(P_0 - 0.32) > 0.20$. The terms in the inequality are replaced necessarily with $m = \bar{x}$ and $P_0 = f_0/N$ from the sample.

Since the number of zero frequencies in Table 1 is relatively large, method 2 is the more promising. As a test of its expected efficiency, we find $(x + 0.17)(f_0/N - 0.32) = .193$, which is very close to the level for 90 percent efficiency. From Eq. 5 the required $\hat{k}_2$ is the trial value k' for which $k' \log(1 + \bar{x}/k') = \log(150/70) = .330993$. The first trial value of $k' = 1$ was based on the estimate from method 1 and the computation arranged as follows:

k'	$1 + \bar{x}/k'$	$k' \log (1 + \bar{x}/k')$
1	2.14667	.331767
.98	2.17007	.329744
.992	2.15592	.330962

With $k'_1 = 1$, the left side of Eq. 5 was too large, so that the calculation was repeated with $k'_2 = .98$, reversing the inequality. By interpolation $k'_3 = .98 + .02\,(.330993 - .329744)/(.331765 - .329744) = .992$, which slightly underestimated the required value. By interpolation between k'_1 and k'_3, $\hat{k}_2 = .99231$.

(3) For many distributions, k cannot be determined by either of the above techniques with an acceptable efficiency. In these and all critical cases, the k computed by Eq. 3 or 5 may be considered as the first step toward a definitive solution. This is provided by the method of maximum likelihood (17, 24). By suitable arrangement of the calculation, as developed by Sir Ronald Fisher in the appendix to the present paper, the procedure is practicable and rapid when the largest observation does not exceed 20 or 30. Scores (z_i) are computed from trial values of k'_i, selected so that they bracket the required estimate $\hat{k}$, for which $z_i = 0$ in the equation

$$z_i = S\left(\frac{A_x}{k'_i + x}\right) - N \ln\left(1 + \frac{\bar{x}}{k'_i}\right) \tag{6}$$

where ln designates a natural logarithm. As a first step in the computation, the accumulated frequency A_x in all units containing more than

x organisms is written opposite each x. The reciprocals $1/(k_i' + x)$ from Barlow's Tables (to seven places or more) are multiplied in turn by A_x and the products accumulated to obtain the summation in the first term. The second term may be determined as the seven-place common logarithm of $(1 + \bar{x}/k_i)$ multiplied by $2.3025851 \times N$.

The first score z_1 $(i = 1)$ is computed with the first trial value k_1', usually based upon the $\hat{k}$ from Eq. 3. The second trial value k_2' depends upon the sign of z_1. If z_1 is positive, $k_2' > k_1'$; if negative, $k_2' < k_1'$, the two differing just enough to give opposite signs to z_1 and z_2. The third trial value k_3', between k_1' and k_2', is obtained by linear interpolation for $z = 0$, but the computed z_3 is rarely exactly zero. For precision, it is preferable to compute a z_4 with a sign opposite to that of z_3 by selecting k_4' at about the same distance as k_3' beyond a newly interpolated k' for $z = 0$. This provides a narrower interval within which the final $\hat{k}$ may be interpolated and a better estimate of its variance obtained.

Since the better of the two approximations of $\hat{k}$ in our example was barely 90 percent efficient, the likelihood solution would be preferred. The cumulative frequencies exceeding each x were first listed (Table 1), so that for $x = 0$, for example, $A_x = 150 - 70 = 80$, and for $x = 1$, $A_x = 80 - 38 = 42$. Starting with a trial value of $k_1' = 1.0$, the reciprocal, $1/(k_1' + x)$, for each x was multiplied by its corresponding A_x and the products accumulated in the calculator to obtain the first term in the score z (Eq. 6). This sum has been listed separately and below it the second term in the score, $345.3878 \times \log(1 + 1.146667/k')$:

	$k_1' = 1.0$	$k_2' = 1.05$	$k_3' = 1.026$	$k_4' = 1.023$
$S\{A_x/(k' + x)\}$	114.9262	110.4045	112.5247	112.7961
$-N \ln (1 + \bar{x}/k')$	−114.5875	−110.7227	−112.5432	−112.7752
z_i	.3387	−.3182	−.0185	.0209

Since z_1 was positive, the second trial k' was increased to $k_2' = 1.05$, leading to a negative z_2. Interpolating between them for $z = 0$, $k_3' = 1.0 + (.3387 \times .05)/(.3387 + .3182) = 1.026$, which, in turn, gave $z_3 = -.0185$. From linear interpolation between z_1 and z_3 for $z = 0$, $\hat{k} = 1.0247$. To insure a positive score near zero, a new trial value was selected of $k_4' = 1.023$. Interpolation between z_3 and z_4 gave the maximum likelihood estimate of $\hat{k} = 1.02459$.

The Variances of $\bar{x}$ and $\hat{k}$. The sampling variances of the statistics of the negative binomial depend upon the parameters m and k, but in practice these are replaced necessarily by the corresponding statistics

$\bar{x}$ and $\hat{k}$. The mean is computed efficiently in all cases and its variance is

$$V(\bar{x}) = \left(m + \frac{m^2}{k}\right)/N \tag{7}$$

The error variance of the mean in the example from Table 1, computed with the maximum likelihood estimate of k, was $V(\bar{x}) = (1.14667 + 1.28330)/150 = .01620$, giving the standard error $s_{\bar{x}} = .1273$.

The variance of $\hat{k}$ depends upon how it has been estimated. In general, the variance of $\hat{k}$ is less when k is small than when it is large.

(1) If computed from s^2 by Eq. 3, its large-sample variance (2) is

$$V(\hat{k}_1) \doteq \frac{2k(k+1)}{NR^2} \tag{8}$$

solved with $R = \bar{x}/(\hat{k} + \bar{x})$. For $\hat{k}_1 = 1.16670$ from Eq. 3, $R = 1.14667/2.31337 = .49567$, and $V(\hat{k}_1) = 5.0558/36.853 = .1372$, from which this estimate of k has a standard error $\sqrt{.1372} = .370$.

(2) If $\hat{k}$ is determined from the number of zero units by Eq. 5, its large-sample variance (2) is

$$V(\hat{k}_2) \doteq \frac{(1-R)^{-k} - 1 - kR}{N[-\ln(1-R) - R]^2} \tag{9}$$

where R is defined as above and ln is a natural logarithm. The error variance in the example for $\hat{k}_2 = .99231$ as estimated by Eq. 5 is solved with $R = 1.14667/2.13898 = .53608$, to obtain $V(\hat{k}_2) = .61089/8.0709 = .07569$. The standard error of $\hat{k}_2$ is 0.2751, or about 3/4 as large as that for $\hat{k}_1$.

(3) The variance of the maximum likelihood estimate of k is the reciprocal of the amount of information about k, or the rate at which the score z is decreasing as it passes the zero. It is computed, therefore, from the two values of z_i just above and below zero, say z_3 and z_4, and the two trial values of k_i' with which they have been computed, k_3' and k_4', as

$$V(\hat{k}) = \frac{k_3' - k_4'}{z_4 - z_3} \tag{10}$$

Hence the error variance of $\hat{k} = 1.02459$ is

$$V(\hat{k}) = (1.026 - 1.023)/(.0209 + .0185) = .07614,$$

so that the maximum likelihood $\hat{k}$ had a standard error of $s_{\hat{k}} = .2759$.

Tests for Agreement with the Negative Binomial. Perhaps the most convincing test of the adequacy of the negative binomial in any given case is the agreement between the frequencies (f) observed at each x and their expected values (ϕ) as computed from the statistics of the sample. The discrepancy between them is easily tested by χ^2.

The expected frequencies are computed with Eq. 1, most readily in succession and starting with the number expected at $x = 0$. This first expectation is determined with the aid of 7-place logarithms as

$$\phi_0 = N/q^k \tag{11}$$

and the succeeding entries for $x = 1, 2, 3, \cdots$ as

$$\phi_x = \frac{(k + x - 1)R}{x} \cdot \phi_{x-1} \tag{12}$$

More decimal places should be retained in the calculator at each stage than need be recorded, so as to avoid accumulating rounding errors. The observed and expected frequencies are then compare by χ^2, where

$$\chi^2 = S\left\{\frac{(f - \phi)^2}{\phi}\right\} \tag{13}$$

χ^2 has three fewer degrees of freedom than the number of ratios that are summed. As usual, the frequencies with small expectations are pooled, preferably so that no expectation is less than 5. If χ^2 shows good agreement between the matched frequencies, no other test may be needed, especially if both $\bar{x}$ and $\hat{k}$ are efficient estimates.

In the example, the expected frequency for $x = 0$ was determined with Eq. 11 as the antilog of log (150) − 1.02459 log (2.11915) or $\phi_0 = 69.4879$, giving the initial entry in the fourth column of Table 1. From $p = 1.11915$ and $q = 1 + p$, $R = 1.11915/2.11915 = .528113$ and by Eq. 12, $\phi_1 = 1.02459 \times .528113 \times 69.4879 = 37.5999$, $\phi_2 = 2.02459 \times .528113 \times 37.5999/2 = 20.1011$ and so on for successive values of x. To avoid rounding errors, six significant figures were carried in the calculation, although the ϕ's were recorded only to two decimal places. The final value, for $x = 8+$, was obtained as the difference between 150 and the sum of preceding ϕ's. The calculation of χ^2 for the discrepancy between the observed and expected frequencies (Eq. 13) is shown in the last column of Table 1, pooling the frequencies for $x \geq 5$ so as to avoid expectations of less than $\phi = 5$. The resulting $\chi^2 = 2.484$ with three degrees of freedom and $P = 0.48$ indicates good agreement with the negative binomial.

The comparison of observed and expected frequencies by χ^2 may be distorted by chance irregularities in the individual entries. Thus in

testing agreement with Neyman's contagious distribution, Beall (6) "smoothed" some of his more uneven observed frequencies before computing χ^2. This difficulty can be avoided by testing the agreement of the observed with the expected second and third moments of the negative binomial. The relation of these tests to alternative formulations, such as the logarithmic, the discrete log-normal, and the "contagious" distributions, has been considered by Anscombe (2). The moment tests have the further advantage that they take account of the few large values which are missed by grouping the tail of an observed distribution in computing χ^2.

Two tests have been described by Anscombe (2), each having the form of a difference between an observed and an "expected" moment. Although m in each test is estimated efficiently by $\bar{x}$, its variance has been derived on the assumption that an efficient estimate of k is not available for computing the expectation. A variance so derived may not apply when the expected moment is estimated by maximum likelihood. Hence the test criteria T and U are defined in terms of the estimates for which the variances are known. These variances, however, should always be computed with the best available estimate of k, usually that derived by maximum likelihood ($\hat{k}$).

The difference T between the third moment of the sample and its value predicted from the first two moments of the same sample of a negative binomial is

$$T = \frac{[x^3]}{N} - s^2\left\{\frac{2s^2}{\bar{x}} - 1\right\} \tag{13}$$

where

$$[x^3] = S\{f(x - \bar{x})^3\} = S(fx^3) - 3\bar{x}S(fx^2) + 2\bar{x}^2S(fx) \tag{14}$$

The significance of the difference T is determined by comparison with its standard error, the square root of its large-sample variance

$$V(T) = 2m(k + 1)p^2q^2[2(3 + 5p) + 3kq]/N \tag{15}$$

The variance of T should be computed with estimates of p, q and k based upon the maximum likelihood $\hat{k}$ when it is known. Although the expected third moment may be determined more accurately from the same maximum likelihood estimates as $q(q + p)m$, the variance in Eq. 15 is then of doubtful applicability.

When the observed second moment is compared with its expectation computed with $\hat{k}_2$, the difference

$$U = s^2 - (\bar{x} + \bar{x}^2/\hat{k}_2) \tag{16}$$

has the large-sample variance

$$V(U) = 2m(k+1)pq^2\left(1 - \frac{R^2}{-\ln(1-R) - R}\right) \Big/ N + p^4 V(\hat{k}_2) \quad (17)$$

$V(\hat{k}_2)$ is defined in Eq. 9 but computed with the maximum likelihood estimate $\hat{k}$ if this is known, as are the other terms in Eq. 17. Here again the expected second moment, qm, can be estimated more exactly with the maximum likelihood value for k but the applicability of the variance in Eq. 17 is then in doubt.

The differences between the observed and expected moments are much easier to compute than their standard errors. From $[x^3] = 778.47$ (Eq. 14) the present example had an observed third moment of 5.1898 and an expected value from the first two moments, $\bar{x}$ and s^2 of 6.7429, so that $T = -1.553$ by Eq. 13. Since its variance by Eq. 15 was 4.1272, the standard error of T, 2.032, showed no discrepancy from a negative binomial. The corresponding difference for the second moment and its error has been computed by Eqs. 16 and 17, to obtain $U = -0.198 \pm 0.302$, again in agreement with the negative binomial.

Models for the Negative Binomial. When the number of individuals per unit of space or time in repeated counts cannot be assumed to have the same expected value, they may represent a mixture of several homogeneous Poisson distributions. The number in each unit is restricted to the integers but this is not true of the expectations or means. In a mixture of Poissons, the means represent a positive continuous variate. The simplest frequency distribution which they might follow is the Eulerian distribution or the Pearson type III curve, and if in fact they are so distributed the observations will conform to the negative binomial.

The expected frequency may be known to vary within an observed distribution. A case in point is the distribution of bacterial clumps over a milk film (20). At least two disturbing factors were involved. In preparing a film, 0.01 milliliter of milk was placed on a microscopic slide and spread with a needle over an area of one square centimeter. Bacteria caught by surface tension on the lower surface of the drop adhered to the glass slide on contact and thus increased the concentration of bacteria in this area of the film. Secondly, the fresh drop did not have the same thickness over the entire square centimeter but due to surface tension was thicker in the center than in the margins, so that more bacteria were deposited in the center. Despite these two factors, milk meeting public health standards had so low a bacterial count that the distribution of bacterial clumps per microscopic field was seldom distinguishable from a Poisson. However, when the

TABLE 2

Observed distributions of bacterial clumps per field (Obs. f) in a milk film (20) and of yeast cells per square in a haemocytometer (27) and the expected frequencies computed from the negative binomial (Bin ϕ) and from the Neyman type A (Ney ϕ) distributions (21).

No. per unit x	Bacterial clumps Obs. f	Bacterial clumps Bin. ϕ	Yeast cells Obs. f	Yeast cells Bin. ϕ	Yeast cells Ney. ϕ
0	56	64.2	213	214.2	214.8
1	104	90.3	128	122.8	121.3
2	80	82.7	37	45.0	45.7
3	62	62.1	18	13.4	13.7
4	42	41.6	3	3.5	3.6
5	27	25.8	1	.9	.8
6	9	15.1		+.2	+.1
7	9	8.5			
8	5	4.7			
9	3	2.5			
10	2	1.3			
11		+1.2			
19	1				
N, $P(\chi^2)$	400	.54	400	.19	.18

bacterial count was high, the distribution of clumps per field reflected this known heterogeneity and then was often a negative binomial, as in the example in Table 2. This phenomenon of substantial agreement with the Poisson at low population densities and with the negative binomial at higher densities has been observed with both plant and animal populations (4, 5). A lack of randomness in microscopic counts was observed by Student in counting yeast cells with a haemocytometer (27). In fact, this seems to be the first case to be fitted with a negative binomial. The original counts are given in Table 2, together with the expected negative binomial frequencies as computed by maximum likelihood and those computed by Neyman (21) with his type A contagious distribution.

The distribution of insect pests is so seldom uniform that most experiments on insect control are randomized and replicated. In a field experiment of this type on the corn borer, four treatments were arranged in 15 randomized blocks (26). At the end of the season, eight hills of corn were selected at random in each plot and the borers recorded from each hill. This experiment has been reported both in

TABLE 3

Distribution of corn borers (Obs. f) in a field experiment arranged in 15 randomized blocks, where treatment 1 is the untreated control or check (6). The expected frequencies for the negative binomial (Bin. ϕ) have been computed independently for each treatment with the statistics in the last row of the table; those expected for the Neyman type A (Ney ϕ) are from Beall (6).

Borers per hill x	Treatment 1			Treatment 2			Treatment 3			Treatment 4		
	Obs. f	Bin. ϕ	Ney. ϕ	Obs. f	Bin. ϕ	Ney. ϕ	Obs. f	Bin. ϕ	Ney. ϕ	Obs. f	Bin. ϕ	Ney. ϕ
0	19	16.6	34.4	24	19.6	22.6	43	44.3	49.8	47	45.3	53.4
1	12	18.5	6.4	16	22.2	16.7	35	31.1	23.3	23	30.1	19.7
2	18	16.9	10.4	16	19.7	18.3	17	19.1	18.9	27	18.4	17.5
3	18	14.5	11.9	18	15.9	16.4	11	11.2	12.3	9	11.0	12.1
4	11	11.9	11.2	15	12.1	13.4	5	6.4	7.3	7	6.4	7.5
5	12	9.5	9.5	9	9.0	10.3	4	3.6	4.1	3	3.7	4.4
6	7	7.5	7.9	6	6.5	7.5	1	2.0	2.2	1	2.1	2.5
7	8	5.9	6.4	5	4.6	5.2	2	1.1	1.1	1	1.2	1.4
8	4	4.5	5.2	3	3.3	3.5	2					
9	4	3.5	4.1	4	2.3	2.3		+1.2	+1.0		+1.8	+1.5
10	1	2.7	3.2	3	1.6	1.5				1		
11		2.0	2.5		1.1	.9				1		
12	1	1.5	1.9	1								
13	1	1.2	1.4		+2.1	+1.4						
15	1											
17	1	+3.3	+3.6									
19	1											
26	1											
N, $P(\chi^2)$	120	.65	.002	120	.66	.98	120	.88	.09	120	.16	.09
$\bar{x}$, k	4.033		1.532	3.167		1.764	1.483		1.333	1.508		1.190

terms of the total number of borers per plot (26) and as frequency distributions showing for each treatment the number of hills with $x = 1, 2, \cdots$ borers (6). From an analysis of variance of the plot totals (in logarithmic units) the level of borer infestation varied significantly from block to block ($P < .01$). In consequence, the composite distributions from the 15 plots for each treatment (Table 3) represented unequal levels of infestation. Negative binomials have been fitted separately to each of them. Since the expected frequencies agreed well with the observed values, the data are consistent with the hypothesis that each represented the sum of several Poisson distributions of unequal means.

The distributions in Table 3 might arise from a different model for the negative binomial in which the non-randomness is attributed to "contagion", in this case, a result of the larvae hatching from eggs that were laid in masses. Contagion, in fact, was the basis for the Neyman type A distribution developed originally for these observations (21, 6) as described in the next section. With a different mathematical formulation it leads also to a negative binomial. "Contagion" was one

TABLE 4

Distribution of the number of accidents experienced by machinists (f) and their negative binomial expectations (ϕ) (16). Distribution of soil bacteria in microscopic counts, showing the colonies per field fitted with the Poisson series, the bacteria per colony with a logarithmic series, and the bacteria per field with a negative binomial (18).

Accidents per machinist	No. of machinists		Colonies per field	No. of fields		Bacteria per colony	No. of colonies		Bacteria per field	No. of fields	
	f	ϕ		Obs.	Calc.		Obs.	Calc.		Obs.	Calc.
0	296	296.7	0	11	14.6	1	359	362.1	0	11	13.0
1	74	71.0	1	37	40.9	2	146	136.1	1	17	21.0
2	26	26.4	2	64	57.2	3	57	68.3	2	31	24.6
3	8	11.0	3	55	53.4	4	41	38.5	3	24	25.4
4	4	4.8	4	37	37.4	5	26	23.2	4	29	24.2
5	4	2.2	5	24	20.9	6	17	14.5	5	18	22.0
6	1	1.0	6+	12	15.6	7+	27	30.3	6	19	19.4
7		.5							7	16	16.7
8	1	.2							8	13	14.1
9		+.2							9	17	11.7
									10	6	9.6
									11	8	7.8
									12+	31	30.5
$N, P(\chi^2)$	414	.57		240	.63		673	.56		240	.52

of the explanations proposed by Student (28), who wrote, "If the presence of one individual in a division increases the chance of other individuals falling into that division, a negative binomial will fit best, but if it decreases the chance, a positive binomial". This explanation has figured prominently in the study of accident statistics (3, 16).

An early example is the distribution in Table 4 of accidents experienced by 414 machinists in three months, where the observed frequencies are matched satisfactorily by those computed with a negative binomial. If each machinist had had the same initial probability of being involved in an accident but if this probability were increased (or decreased) by his having an accident, contagion would be present and a negative binomial distribution could result. However, an opposite assumption leads to exactly the same expected distribution. If experiencing an accident had no effect upon the risk of another accident, but if the individual machinists or their shops or intervals within the three month period differed in their accident-proneness, a negative binomial would also result. Hence, the appearance of "contagion is not inherent in nature but simply in our method of sampling" (11). The relation of these two models, a mixed or compound Poisson distribution without contagion and contagion which changes the odds of further events, is discussed ably in the recent monograph by Arbous and Kerrich (3). Alternative distributions have been developed from other math-

TABLE 5

Observed distributions of quadrat counts (Obs. f) of *Lespedeza capitata* and of *Liatris aspera* in an old field association (30) and of *Primula auricula* in a grassland association (7), and their expectations with the negative binomial (Bin. ϕ), Neyman contagious type A (Ney ϕ) and Thomas double Poisson (Thom ϕ) distributions.

Plants per quadrat x	Lespedeza capitata Obs. f	Bin. ϕ	Ney* ϕ	Thom ϕ	Liatris aspera Obs. f	Bin ϕ	Thom ϕ	Primula Obs. f	Bin ϕ
0	7178	7178.1	7188.4	7279.2	7403	7403.1	7420.4	26	23.6
1	286	283.7	219.6	105.2	183	179.8	140.0	21	26.1
2	93	95.0	140.8	127.9	34	40.0	62.3	23	20.9
3	40	41.1	61.6	78.6	14	11.5	14.4	14	14.7
4	24	19.8	21.1	33.1	4	3.7	2.4	11	9.5
5	7	10.2	6.2	11.1	1	1.3	0.4	4	5.9
6	5	5.4	1.7	3.4	1	.4	0.1	5	3.5
7	1	2.9	.5	1.0		+.2		4	2.1
8	2	1.6	.1	+.5					1.2
9	1	.9						1	.7
10	2	.5							+.8
11	1	.3							
12		+.5							
N, $P(\chi^2)$	7640	.84	<.001	<.001	7640	.46	<.001	109	.70

*Computed by maximum likelihood (30), all other Neyman Type A expectations fitted by moments.

ematical definitions of "contagion" and will be considered in the next section.

The negative binomial may result from a different but related model. In counts of soil bacteria, Jones and Mollison (18) recorded for each microscopic field both the number of bacterial colonies and the number of bacteria in each colony. The number of colonies per field agreed well with the Poisson expectation for a random distribution and the number of bacteria per colony in the same counts with Fisher's logarithmic distribution. Under these conditions, the expected distribution of bacteria per field is a negative binomial (23), as confirmed by the agreement of the expected and observed frequencies (Table 4).

Quadrat counts in plant ecology which departed from Poisson have been attributed to the occurrence of plants in "clumps". Blackman (7) noted that *Primula auricula* reproduces vegetatively by short rhizomes, so that older individuals are often surrounded by younger plants. In an old-field community (30) both *Liatris astera* and *Lespedeza capitata* tended to occur in clumps. The distributions of clumps per quadrat and of plants per clump have not been reported separately, so that the model cannot be tested directly. However, the distribution of plants per quadrat in all three cases agreed excellently with the negative binomial (Table 5).

TABLE 6

Observed animal distributions (f) and their negative binomial expectations (ϕ) of *Microcalanus* nauplii in samples of marine plankton (also fitted with a Neyman type *A*) (5), of *Tanytarsus* in Ekman hauls (19), of Oligochaetes in Petersen hauls (19), of isopods under boards (9), and of the mite *Liponysus bacoti* on rats in Savannah (10).

Individuals per unit x	Microcalanus		Ney	Tanytarsus		Oligochaetes		Isopods		Mites	
	f	ϕ	ϕ	f	ϕ	f	ϕ	f	ϕ	f	ϕ
0		.1	.8	32	29.5	39	34.7	28	30.2	160	160.0
1	2	.8	1.9	28	32.5	24	29.6	28	21.6	19	15.9
2	4	2.1	3.7	25	29.0	18	23.6	14	16.2	11	8.5
3	3	4.3	5.8	34	23.9	21	18.3	11	12.3	6	5.8
4	5	7.1	8.0	13	18.9	15	14.1	8	9.4	5	4.3
5	8	9.9	10.0	14	14.5	15	10.7	11	7.3	4	3.4
6	16	12.4	11.6	17	10.9	6	8.2	2	5.6	4	2.8
7	13	14.0	12.5	5	8.1	8	6.2	3	4.3	3	2.4
8	12	14.7	12.9	6	6.0	6	4.6	3	3.4	2	2.0
9	13	14.5	12.7	1	4.4	2	3.5	3	2.6	2	1.8
10	15	13.5	11.9	9	3.2	1	2.6	3	2.0		1.6
11	15	12.0	10.9		2.3	2	2.0	2	1.6		1.4
12	9	10.3	9.6		1.7	3	1.5		1.2	1	1.2
13	9	8.5	8.2	2	1.2	3	1.1	1	.9		1.1
14	7	6.8	6.8	1	.9		.8	2	.7		1.0
15	4	5.2	5.5		.6	1	.6	1	.6	2	.9
16	4	4.0	4.4		.4		+1.9		.4		.8
17	6	2.9	3.4	1	.3			2	.3		.8
18	2	2.1	2.6	1	.2				+1.4	1	.7
19		1.5	1.9		+.5					1	.6
20	2	1.1	1.4							6	+10.0
21	1	.7									
22		+1.5	+3.5								
N, $P(\chi^2)$	150	.89	.64	189	.14	164	.53	122	.34	227	.59

Other models have been developed for the negative binomial (2) and there is no reason to suppose that the possibilities have been exhausted. This is suggested in part by the variety of its applications to biological data. Some applications to fresh-water dredge samples (19) and to marine plankton (5) are shown in Table 6. It has formed the basis for a sequential sampling scheme for tapeworm cysts in whitefish (22). It has described effectively the distribution of insects in the field, including the beet leafhopper (8) and the wireworm (31), of ticks on individual sheep (12), of mites on rats (10) and of isopods under boards (9) (Table 6). Some failures in fitting can be ascribed to the inefficiency of the moment estimate of k. One of these is a count of *Ribes* on Mt. Spokane (14, 31), where Equation 3 gave $\hat{k}_1 = .134$ and Equation 6 gave $\hat{k} = .205$. Even though the estimates did not differ significantly,

TABLE 7

Distributions of quadrat counts with apparent bimodality (Obs. f) and their expected frequencies for the negative binomial (Bin ϕ), Neyman type A (Ney ϕ) and Thomas double Poisson (Thom ϕ) distributions, representing three species in a salt marsh, *Salicornia stricta*, *Plantago maritima* and *Ameria maritima* (4, 29), and a weed on arable land, *Chenopodium album* (25).

Plants per quadrat x	Salicornia Obs. f	Salicornia Bin ϕ	Salicornia Ney ϕ	Plantago Obs. f	Plantago Bin ϕ	Plantago Thom ϕ	Ameria Obs. f	Ameria Bin ϕ	Ameria Ney ϕ	Ameria Thom ϕ	Chenopodium Obs. f	Chenopodium Bin ϕ	Chenopodium Thom ϕ
0	4	3.3	10.7	12	7.6	11.0	57	54.1	54.9	56.4	19	9.2	19.0
1	3	6.4	4.0	8	11.3	6.7	6	16.2	7.9	5.6	5	13.5	5.0
2	8	8.4	6.5	9	12.4	10.7	12	9.0	10.1	10.0	6	14.3	9.7
3	13	9.4	7.8	13	12.0	11.2	5	5.8	9.0	9.6	9	13.0	10.6
4	11	9.6	8.1	6	10.8	10.8	5	3.9	6.5	6.8	5	11.0	9.6
5	9	9.2	7.9	8	9.3	10.0	5	2.8	4.3	4.3	20	8.8	8.3
6	8	8.4	7.6	11	7.8	8.8	7	2.0	2.8	2.8	14	6.8	7.2
7	10	7.5	7.0	7	6.4	7.4	1	1.5	1.8	1.8	8	5.2	6.0
8	3	6.5	6.4	8	5.1	6.0		1.1	1.1	1.1	4	3.8	4.9
9	3	5.6	5.7	7	4.0	4.7	1	.8	.9	.7	3	2.8	3.8
10	8	4.7	4.9	3	3.2	3.6	1	.6	.4	.4	2	2.0	3.0
11	3	3.1	4.2	4	2.5	2.7						+4.6	+7.9
12	4	2.5	3.5	1	1.9	2.0		+2.2	+.3	+.5			
13	4	2.1	2.9	1	1.4	1.4							
14		1.7	2.4										
15	3	1.3	1.9										
16		1.0	1.5	1	+4.3	+3.0							
17		.8	1.2										
18	1	.6	.9										
19				1									
20+	3	+5.9	+2.9										
N, $P(\chi^2)$	98	.48	.17	100	.14	.74	100	.014	.53	.29	95	<.001	<.001

the χ^2 test for the agreement of the observed and expected frequencies from the two estimates gave $P = .017$ and $P = .25$ respectively.

Solely as a method for summarizing a set of observations with two statistics, one of them the mean, the negative binomial should be of increasing interest to biologists. An adequate fit of this distribution to the data may serve to justify further statistical analysis such as sequential sampling (22) or a transformation for stabilizing the variance preparatory to the analysis of variance. But since several quite different models might possibly underlie data which conform to the negative binomial, one cannot use this agreement as the sole basis for justifying a particular model or conclusions based upon it.

Comparisons with Other Distributions for Over-Dispersion. Although the negative binomial is the easiest to compute and the most widely applicable of the distributions for over-dispersion, several others have been proposed. Some of these have two or more modes, while the negative binomial has only a single mode. In fitting the negative binomial to observed distributions, any "extra" modes are assumed to represent

random variation, as in fitting the negative binomial to the distributions of corn borer for treatments 1 and 2 in Table 3, of *Tanytarsus* and of *Microcalanus* in Table 6, and of quadrat counts in Table 7. In some cases, this assumption worked well, in others passably and in a few cases badly. Three two-parameter distributions have been described which may have two or more modes, the Neyman contagious type *A*, the Thomas double Poisson and the Polya. Each has been based upon a mathematical model of biological interest.

The Neyman contagious type *A* distribution (21) assumes an initial population in which groups are dispersed uniformly, within the limits of the Poisson, over the area (or period) represented by the final counts. Individuals then move out from these random centers independently but at too slow a rate to equalize their dispersion over the entire area. As numerical examples, Neyman cites the distribution of corn borers following treatment 2 (Table 3) and Student's haemocytometer counts of yeast (Table 2). In accord with theory, corn borer eggs are laid in masses and the larvae on hatching tend to migrate to neighboring corn plants. The Neyman expected frequencies, as fitted by Beall, are shown in Table 3. For treatment 2 they reproduced the observations better than the negative binomial but the reverse was true for the remaining three treatments. In the *Armeria* counts of Table 7, the Neyman type *A* again reproduced the bimodality which was missed by the negative binomial, but when fitted to unimodal distributions, the negative binomial did as well or better (Tables 2, 5, 6, 7). Fracker and Brischle (14) fitted the Neyman type *A* curve to six series of *Ribes* counts. None of them approached agreement but five of the six agreed well with the negative binomial when computed by method 1 by Wadley (31) and the sixth agreed when fitted efficiently as noted above. Despite the advantage of a potentially multimodal curve, the range of distributions fitted by the Neyman curve seems to be more restricted than with the negative binomial.

The Thomas double Poisson distribution (29) is also potentially multimodel with peaks that are somewhat more sharply defined than in the Neyman type *A*. It assumes that the number of plants per quadrat can be broken down into two Poissons, one of the number of cluster centers and the other of the number of additional plants (after the first) in a cluster. Thus individual plants of *Plantago maritima* tended to be grouped, possibly because their inflorescences are short compact spikes which frequently fall to the ground with the seeds still in the capsules (4). Frequencies computed by moments for this and other series with the Thomas distribution are given in Tables 5 and 7. They closely

resemble the expectations for the Neyman type A and have similar advantages and limitations.

Under certain conditions, the Polya distribution (2) may have two modes with somewhat larger frequencies at $x = 0$ than in the negative binomial, which in other respects it closely parallels. Under certain conditions, the Polya distribution might represent the number of individuals per quadrat in a growing population better than the negative binomial (2). If, for example, the original progeniters were released all at the same time rather than continuously over a period, the Polya distribution would be indicated, but if the individual rates of birth and death were constant and immigration occurred at a constant rate, the negative binomial would be more appropriate. So far as most biological distributions are concerned, the Polya seems to be a minor variant of the negative binomial.

A quite different alternative is provided by Fisher's logarithmic distribution (13). In studies of species, area and abundance, this has been applied effectively to the number of species (f) represented by $x = 1, 2, 3, \cdots$ individuals, in the catches of insect light traps for example (33). If k in a negative binomial approaches zero and we omit the zero class, the observed frequencies can be considered a sample from a logarithmic distribution. Williams (33) has fitted the logarithmic distribution to Buxton's data on the number of Hindu male prisoners in a south Indian jail with $x = 1, 2, 3, \cdots$ lice per head, omitting the 612 individuals (see Anscombe's correction, 2) who were free of lice. The observed frequencies agreed far better with the logarithmic expectations than with those computed from the negative binomial by method 1. In this instance, however, method 1 has an efficiency of less than 50 percent and method 2 and efficiency of nearly 98 percent. When recomputed with $\bar{x} = 6.9357 \pm .5634$ and $\hat{k}_2 = .144198 \pm .008195$, the divergence between the expected and observed frequencies (Table 8) was well within the sampling error ($\chi^2 = 31.74$, $n = 32$). Nevertheless, the observed second and third moments were significantly larger than their expectations, with $U = 243.3 \pm 38.9$ and $T = 26488 \pm 2048$. Yet, $\hat{k}_2$ was significantly larger than zero, its expected value for a logarithmic distribution.

Reexamination of the data showed that four of the 1,073 prisoners had more than 200 head lice. When $\hat{k}_2$ was recomputed without these individuals, the second and third moments were no longer discrepant, with $U = 22.86 \pm 25.68$ and $T = 575 \pm 3297$. This example indicates the disproportionate effect upon T and U of a very few large values of x. With these omissions, the expected frequencies agreed still more closely

TABLE 8

The observed distribution (f) of lice of all stages on the heads of Hindu male prisoners in Cannamore, South India, in 1937–39, and the frequencies computed with the negative binomial (Bin) from all of the data (ϕ), and from all except four prisoners having more than 200 lice (ϕ'), compared with the frequencies computed for Fisher's logarithmic distribution (Log ϕ) by omitting the 612 prisoners without lice (33, 2).

Lice per head x	No. of heads				Lice per head x	No. of heads			Lice per head x	No. of heads		
	Obs.	Binomial		Log		Obs.	Bin	Log		Obs.	Bin	Log
	f	ϕ	ϕ'	ϕ		f	ϕ	ϕ		f	ϕ	ϕ
0	612	612.0	612.0		11	3	9.6	8.4	22–23	8	8.3	7.0
1	106	86.5	90.5	107.2	12	10	8.7	7.6	24–25	7	7.4	6.2
2	50	48.5	50.8	52.8	13	8	8.0	6.9	26–27	10	6.6	5.6
3	29	33.9	35.5	34.7	14	6	7.4	6.3	28–29	6	6.0	5.0
4	33	26.1	27.3	25.6	15	3	6.8	5.8	30–32	2	7.9	6.7
5	20	21.2	22.1	22.2	16	6	6.3	5.4	33–35	7	6.9	5.9
6	14	17.8	18.5	16.6	17	7	5.9	5.0	36–38	9	6.0	5.2
7	12	15.3	15.8	14.0	18	4	5.5	4.6	39–41	5	5.3	4.6
8	18	13.4	13.8	12.1	19	7	5.1	4.3	42–45	5	6.1	5.3
9	11	11.9	12.2	10.6	20	7	4.8	4.1	46–49	7	5.2	4.6
10	11	10.6	10.9	9.4	21	3	4.5	3.8	50+	27	37.5	37.5

with the observed values, as evidenced by the expectations for $x = 0$ to 10 in Table 8. The agreement of the negative binomial expectations with the observed frequencies of mites on individual rats in Table 6 and of ticks on sheep as given by Fisher (12) suggest a greater usefulness for the negative binomial than for the logarithmic distribution in studies on ectoparasites.

Fitting a Single k to Several Negative Binomial Distributions. The analysis and interpretation of most biological data are facilitated by stability in the variance, so that only means need to be compared. A similar stability in the coefficient k would both increase the utility of the negative binomial and increase our confidence in its suitability for a given problem. The assumption of stability is essential in some cases as in the sequential sampling of whitefish for tapeworm cysts (22). By a simple expansion of the maximum likelihood solution, a combined $\hat{k}_c$ can be computed from a series of distributions and their homogeneity in respect to k tested by χ^2.

The calculation consists of computing the score z for each component distribution with the same trial values of k' and adding the scores for each k' over all component distributions. Trial values are selected until two sums of the scores, $S(z)$, are obtained which closely bracket 0. By interpolation between them, the required estimate $\hat{k}_c$ is that for which $S(z) = 0$. If these sums are designated as z_3 and z_4 from corresponding

trial values of k_3' and k_4', the error variance of $\hat{k}_c$ may be computed by equation 10.

The homogeneity of the k's in the component distributions depends upon the z_3 and z_4 in each individual series for k_3' and k_4'. The ratio

$$z_3^2(k_4' - k_3')/(z_3 - z_4) \qquad (18)$$

is computed from each component distribution and from the totals over

TABLE 9

Calculation of a combined $\hat{k}_c$ by maximum likelihood from the four distributions of corn borer in Table 3. $N \ln(10) = 276.310212$.

Treatment No.	Term	Calculation of score with Eq. 6 for $k_1' = 1.5$	$k_2' = 1.4$	$k_3' = 1.47$	$k_4' = 1.473$	$.003\, z_3^2 / (z_3 - z_4)$
1	$S\{A_x/(k' + x)\}$	156.6745	164.1457	158.8287	158.6100	
	$N \ln(1 + \bar{x}/k')$	156.6387	162.7297	158.4110	158.2317	
	z_i	.0358	1.4160	.4177	.3783	.0133
2	$S\{A_x/(k' + x)\}$	138.3464	145.2368	140.3318	140.1302	
	$N \ln(1 + \bar{x}/k')$	136.1976	141.8773	137.8480	137.6810	
	z_i	2.1488	3.3595	2.4838	2.4492	.5349
3	$S\{A_x/(k' + x)\}$	81.5615	86.2613	82.9113	82.7742	
	$N \ln(1 + \bar{x}/k')$	82.5085	86.6970	83.7206	83.5979	
	z_i	−.9470	−.4357	−.8093	−.8237	.1365
4	$S\{A_x/(k' + x)\}$	81.3606	85.9803	82.6881	82.5532	
	$N \ln(1 + \bar{x}/k')$	83.5105	87.7329	84.7322	84.6083	
	z_i	−2.1499	−1.7526	−2.0441	−2.0551	1.1395
Total	Ratios $S(z_i)$	−.9123	2.5872	.0481	−.0513	1.8242 .0001

$\chi^2 = 1.8241$

By interpolation $\hat{k}_c = 1.47145$, $V(\hat{k}_c) = .003/(.0481 + .0513) = .03018$ by Eq. 10.

all distributions. The sum of the ratios computed from the g individual distributions for k_3' and k_4' is diminished by the ratio computed from the corresponding totals of z_3 and z_4. The difference is χ^2 for testing the homogeneity of k with $g - 1$ degrees of freedom. In solving this expression, it is immaterial which boundary value, z_3 or z_4, is used, since the same χ^2 is obtained with either one.

The procedure can be illustrated with the four distributions of corn borer in Table 3, representing four different treatments in the same

field. The calculation requires for each treatment the mean $\bar{x}$ and the accumulated frequencies A_x corresponding to those in the third column of Table 1. Starting with a trial value of $k_1' = 1.5$, the calculations in Table 9 gave $S(z) = -.9123$, so that a smaller trial value, $k_2' = 1.4$, was selected next, giving $S(z) = 2.5872$. By interpolation between k_1' and k_2' for $S(z) = 0$, $k_3' = 1.47$ and from the resulting $S(z)$ by further interpolation, $k_4' = 1.473$. Interpolation between k_3' and k_4' gave the maximum likelihood estimate $\hat{k}_c$. Testing the homogeneity of k over the four treatments required the ratios in the last column of Table 9, from which $\chi^2 = 1.824$ with three degrees of freedom. We conclude that the k's for the four treatments did not differ significantly and could be represented by a single value of $\hat{k}_c = 1.4715 \pm .1737$.

Summary. In analyzing biological counts for which the variance is significantly larger than the mean, the value of the negative binomial distribution is enhanced by a simplified maximum likelihood method for estimating the coefficient k. The calculation is illustrated in detail with a numerical example and compared with other estimates of the same parameter. The error variance of the statistics of the negative binomial and tests for its agreement with a set of observations are illustrated with the same example. Models underlying the negative binomial are reviewed with reference to observed distributions of both plants and animals. A comparison with other distributions for over-dispersion suggests that the negative binomial is the most widely adaptable and generally useful of those that have been proposed so far. Finally, the maximum likelihood estimation is extended to the calculation of a single coefficient k from a series of similar distributions and the testing of their homogeneity by χ^2.

Acknowledgements. I am indebted especially to Sir Ronald Fisher, not only for the method upon which this paper depends so largely, but also for his generous guidance in applying it to many observed distributions. Dr. Philip Garman of The Connecticut Agricultural Experiment Station and Dr. E. S. Deevey, Jr. of Yale University have kindly supplied me with original data from their files, and I acknowledge with thanks the aid of Miss Theresa Santilli and Miss Margaret Robertson with the calculations.

Note on the Efficient Fitting of the Negative Binomial

R. A. Fisher

When it is desired to examine the representation of data having a_x counts of x, for values of x from 0 upward, by means of the negative binomial distribution, in which the expectation of a_x

$$E(a_x) = N \frac{(k + x - 1)!}{x!(k - 1)!} \cdot \frac{p^x}{(1 + p)^{k+x}}$$

is expressed in terms of two parameters p and k, it is well known that the equation of estimation based on the mean

$$pk = \bar{x}$$

is fully efficient.

A second equation, with efficiency varying with the circumstances, may be taken from the second moment or variance

$$p(p + 1)k = s^2$$

or, among other ways, from the frequency of zeros

$$(1 + p)^k = N/a_0$$

In 1941, the author gave a number of rules (12, p. 185) for judging when the first of these is of adequate efficiency, and in 1950 (2), Anscombe has examined more fully the conditions of efficiency of both of these approaches. Many, however, will wish to use these methods only as a first step towards a fully efficient fitting, and the procedure for doing this, whatever means are used for a first orientation, is perhaps worth setting out.

Efficient scoring for k. From the primary expectation

$$m_x = E(a_x) = N \frac{(k + x - 1)!}{x!(k - 1)!} \cdot \frac{p^x}{(1 + p)^{k+x}}$$

we have (using natural logarithms throughout)

$$\frac{\partial}{\partial p} (\log m_x) = \frac{x}{p} - \frac{k + x}{1 + p}$$

whence

$$S\left\{a_x \frac{\partial}{\partial p} (\log m_x)\right\} = \frac{1}{p(1+p)} S(xa_x) - \frac{k}{1+p} S(a_x)$$

$$= \frac{N}{p(1+p)} (\bar{x} - pk)$$

If, therefore, we choose p such that

$$p = \bar{x}/k$$

the likelihood will be maximized for variation of p.

The second equation for maximum likelihood is derived from

$$\frac{\partial}{\partial k} (\log m_x) = F(k + x - 1) - F(k - 1) - \log (1 + p)$$

where $F(z)$ stands for

$$\frac{d}{dz} \log (z!)$$

and

$$F(z) - F(z - 1) = 1/z.$$

The efficient score for k is therefore

$$S\left\{a_x \frac{\partial}{\partial x} (\log m_x)\right\}$$

$$= S\left\{a_x\left(\frac{1}{k} + \frac{1}{k+1} + \cdots + \frac{1}{k+x-1}\right)\right\} - N \log \left(1 + \frac{\bar{x}}{k}\right)$$

In calculating the numerical value of this score for any trial value k, it is convenient first to add up the series of observations from the highest value backward, so that A_x is the number of observations exceeding x, i.e.

$$A_x = a_{x+1} + a_{x+2} + \cdots \text{ ad inf.}$$

Then the convenient expression for the score is

$$S\left(\frac{A_x}{k+x}\right) - N \log \left(1 + \frac{\bar{x}}{k}\right)$$

Trial values are then not difficult to evaluate. The value of k having maximum likelihood is $\hat{k}$, that for which the score vanishes; the corresponding value for p is $\bar{x}/\hat{k}$, and the amount of information about k is,

as usual, the rate at which the score is decreasing as it passes the zero. Hence, the sampling variance and the standard deviation of the estimate may be calculated (p. 182).

BIBLIOGRAPHY

(1) Anscombe, F. J. The statistical analysis of insect counts based on the negative binomial distribution. *Biometrics 5:* 165–173, 1949.

(2) Anscombe, F. J. Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika 37:* 358–382, 1950.

(3) Arbous, A. G. and Kerrich, J. E. Accident statistics and the concept of accident proneness. *Biometrics 7:* 340–432, 1951.

(4) Archibald, E. E. A. Plant populations. I. A new application of Neyman's contagious distribution. *Ann. Bot. 12:* 221–235, 1948.

(5) Barnes, H. and Marshall, S. M. On the variability of replicate plankton samples and some applications of "contagious" series to the statistical distribution of catches over restricted periods. *J. Marine Biol. Assoc. U. K. 30:* 233–263, 1951.

(6) Beall, G. The fit and significance of contagious distributions when applied to observations on larval insects. *Ecology 21:* 460–474, 1940.

(7) Blackman, G. E. A study by statistical methods of the distribution of species in grassland associations. *Ann. Bot. 49:* 749–777, 1935.

(8) Bowen, M. F. Population distribution of the beet leafhopper in relation to experimental field-plot lay-out. *J. Agr. Research 75:* 259–278, 1947.

(9) Cole, L. C. A theory for analyzing contagiously distributed populations. *Ecology 27:* 329–341, 1946.

(10) Cole, L. C. The measurement of interspecific association. *Ecology 30:* 411–424, 1949.

(11) Feller, W. On a general class of "contagious" distributions. *Ann. Math. Stat. 14:* 389–400, 1943.

(12) Fisher, R. A. The negative binomial distribution. *Ann. Eugenics 11:* 182–187, 1941.

(13) Fisher, R. A., Corbett, A. S. and Williams, C. B. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Animal Ecology 12:* 42–58, 1943.

(14) Fracker, S. B. and Brischle, H. A. Measuring the local distribution of ribes. *Ecology 25:* 283–303, 1944.

(15) Garman, Philip. Original data on European red mite on apple leaves. Connecticut, 1951.

(16) Greenwood, M. and Yule, G. U. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *J. Roy. Stat. Soc. 83:* 255–279, 1920.

(17) Haldane, J. B. S. The fitting of binomial distributions. *Ann. Eugenics 11:* 179–181, 1941.

(18) Jones, P. C. T., Mollison, J. E. and Quenouille, M. H. A technique for the quantitative estimation of soil microorganisms. Statistical note. *J. Gen. Microbiology 2:* 54–69, 1948.

(19) Juday, C. Unpublished data on the macroscopic fresh-water fauna in dredge samples from the bottom of Weber Lake, 1942. Courtesy of E. S. Deevey, Jr.

(20) Morgan, M. E., MacLeod, P., Anderson, E. O. and Bliss, C. I. A sequential procedure for grading milk by microscopic counts. *Storrs Agr. Expt. Sta. Bull. 276*, 1951.

(21) Neyman, J. On a new class of "contagious" distributions, applicable in entomology and bacteriology. *Ann. Math. Stat. 10:* 35–57, 1939.

(22) Oakland, G. B. An application of sequential analysis to whitefish sampling. *Biometrics 6:* 59–67, 1950.

(23) Quenouille, M. H. A relation between the logarithmic, Poisson and negative binomial series. *Biometrics 5:* 162–164, 1949.

(24) Sichel, H. J. The estimation of the parameters of a negative binomial distribution with special reference to psychological data. *Psychometrika 16:* 107–127, 1951.

(25) Singh, B. N. and Chalam, G. V. A quantitative analysis of the weed flora on arable land. *J. Ecol. 25:* 213–221, 1937.

(26) Stirrett, G. M., Beall, G. and Timonin, M. A field experiment on the control of the European corn borer, *Pyrausta nubilalis* Hubn. by *Beauveria bassiana* Vuill. *Scient. Agric. 17:* 587–591, 1937.

(27) Student. On the error of counting with a haemocytometer. *Biometrika 5:* 351–360, 1907.

(28) Student. An explanation of deviations from Poisson's law in practice. *Biometrika 12:* 211–215, 1919.

(29) Thomas, M. A generalization of Poisson's binomial limit for use in ecology. *Biometrika 36:* 18–25, 1949.

(30) Thomson, G. W. Measures of plant aggregation based on contagious distribution. *Contr. Lab. Vert. Biol. U. of Mich. No. 53*, 1952.

(31) Wadley, F. M. Notes on the form of distribution of insect and plant populations. *Ann. Ent. Soc. Am. 43:* 581–586, 1950.

(32) Whitaker, L. On the Poisson law of small numbers. *Biometrika 10:* 36–71, 1914.

(33) Williams, C. B. Some applications of the logarithmic series and the index of diversity to ecological problems. *J. Ecology 32:* 1–44, 1944.

THE FITTING OF MULTI-HIT SURVIVAL CURVES

A. W. Kimball

Oak Ridge National Laboratory

1. *Introduction.*

Because of the rapid growth of research in the field of atomic energy for both military and civilian uses, biologists are placing more and more emphasis on the study of the effects of radiation on living organisms. The use of microorganisms in radiation studies has increased steadily since they are in general easier and cheaper to work with than higher forms of life and since the results of such research may be used to define areas of investigation with more expensive material. For the most part data from experiments with microorganisms are survival proportions of cultures exposed to different experimental conditions, frequently varying amounts or intensities of radiation. For this reason the interpretation of survival curves plays a very important role in the field of modern radiobiology.

Among research workers who have realized this fact are Atwood and Norman (1949) who have discussed rather thoroughly the theoretical approach to survival curves from several different points of view. Graphical methods of estimating parameters have been suggested by these and other authors, but the problem of estimating sampling errors seems to have been overlooked. The purpose of this paper is to present some statistical methods for obtaining parameter estimates and their standard errors. Although the analysis of a single-hit curve is simple and straightforward, it will be presented first for completeness.

2. *The single-hit curve.*

If a population consists of organisms each having one sensitive unit inactivation of which causes loss of viability, the probability that the unit is not hit when exposed to a dose x of radiation is assumed to be e^{-kx} where k is some positive constant independent of dose. It follows, therefore, that the expected proportion of a population surviving a dose x is equal to e^{-kx}. If several cultures of the same population are exposed to different doses x_i $(i = 1, \cdots, p)$ and the observed proportions surviving are S_i, the equation*

$$y_i = \log S_i = -kx_i + \epsilon_i , \tag{1}$$

*Throughout the paper all logarithms are taken to the base e.

where ϵ_i represents the amount by which y_i differs from its expected value, provides a simple method for estimating k. If the ϵ_i satisfy certain assumptions (see section 4), an estimate of k is given by

$$\hat{k} = -\sum_{i=1}^{p} x_i y_i / \sum_{i=1}^{p} x_i^2 ,$$

and the variance of $\hat{k}$ is estimated by*

$$s_{\hat{k}}^2 = s^2 / \sum x^2,$$

where

$$s^2 = \frac{1}{p-1}(\sum y^2 + \hat{k} \sum xy).$$

TABLE 1

FITTING OF A SINGLE HIT CURVE, EQ. (1)

[Data from Lea, Haines and Coulson (1936)]

Proportion surviving	Log S	Dose in minutes of exposure
(S)	(y)	(x)
.90	−0.105	1.7
.84	−0.174	2.5
.80	−0.223	4.9
.69	−0.371	9.8
.58	−0.545	14.7
.47	−0.755	27.7
.18	−1.715	55.6
.06	−2.813	85.7

$$\sum y^2 = 11.9499, \qquad \sum x^2 = 11{,}548, \qquad \sum xy = -370.707$$

$$\hat{k} = -(-370.707)/11{,}548 = .0320992$$

$$s^2 = \tfrac{1}{7}[11.9499 + .0320992(-370.707)]$$

$$= .007271$$

$$s_{\hat{k}}^2 = .007271/11{,}548 = .630 \times 10^{-6}$$

$$s_{\hat{k}} = .0008$$

*Indices on summation symbols and subscripts on variables have been omitted wherever the meaning is unambiguous.

In Table 1 the method is illustrated with data from Lea, Haines and Coulson (1936). Spores of *B. mesentericus* were exposed to the beta rays of a radon source for varying lengths of time. Different numbers of loops were exposed at each dose in an attempt to keep the errors uniform. The estimate $\hat{k} = .0321$ when expressed in reciprocal seconds is 5.35×10^{-4} which agrees well with the value 5.25×10^{-4} given by the authors.

3. *The multi-hit curve.*

Frequently populations consist of organisms each having n sensitive units all of which must be inactivated before the organism will lose its viability. If the hits are independent and if k is the same for each unit, the probability that all n units are inactivated is $(1 - e^{-kx})^n$. Accordingly, the expected proportion of a population surviving a dose x is $1 - (1 - e^{-kx})^n$.

One method for fitting survival curves based on this model has been used widely. For large doses

$$(1 - e^{-kx})^n \sim (1 - ne^{-kx}) \tag{2}$$

so that in an experiment resulting in observed proportions surviving S_i over a range of doses x_i $(i = 1, \cdots, p)$, the equation

$$y_i = \log S_i = \log n - kx_i + \epsilon_i \tag{3}$$

leads to a method for estimating n and k. In practice the data are plotted on semi-logarithmic paper and only those points at large doses which appear to lie nearly on a straight line are used in fitting (3). If the assumptions about ϵ_i are correct, estimates of k and n are given by the familiar formulas

$$\hat{k} = -\sum (x - \bar{x})(y - \bar{y}) / \sum (x - \bar{x})^2$$

$$\log \hat{n} = \bar{y} + \hat{k}\bar{x}$$

where $\bar{x}$ and $\bar{y}$ are arithmetic means. The variances of $\hat{k}$ and $\log \hat{n}$ are

$$s_{\hat{k}}^2 = s^2 / \sum (x - \bar{x})^2$$

$$s_{\log \hat{n}}^2 = s^2 \sum x^2 / p \sum (x - \bar{x})^2$$

where

$$s^2 = \frac{1}{p - 2} [\sum (y - \bar{y})^2 + \hat{k} \sum (x - \bar{x})(y - \bar{y})].$$

TABLE 2

FITTING OF A MULTI-HIT CURVE, EQ. (3)

[Data from Pomper (1952)]

Proportion surviving	Log S	X-ray dose in 10^4 roentgens
(S)	(y)	(x)
.86		1
.64		2
.....	.....	.....
.21	−1.561	4
.072	−2.631	6
.020	−3.912	8
.0056	−5.185	10
.0020	−6.215	12
.00045	−7.706	14

$$\sum (x - \bar{x})^2 = 70.00 \qquad \bar{x} = 9.00$$

$$\sum (y - \bar{y})^2 = 26.158162 \qquad \bar{y} = -4.535$$

$$\sum (x - \bar{x})(y - \bar{y}) = -42.75 \qquad \sum x^2 = 556$$

$$\hat{k} = -(-42.75/70) = .610714$$

$$\log \hat{n} = -4.535 + .610714(9.00) = .9614$$

$$\hat{n} = 2.62$$

$$s^2 = \tfrac{1}{4}[26.158162 + (.610714)(-42.75)]$$

$$= .01254$$

$$s^2_{\log \hat{n}} = (.01254)(556)/6(70) = .0166$$

$$s_{\log \hat{n}} = .13$$

$$s^2_{\hat{k}} = (.01254)/70 = .000179$$

$$s_{\hat{k}} = .0134$$

This method is illustrated in Table 2 with some hitherto unpublished data of Dr. S. Pomper of Oak Ridge National Laboratory. These survival data are from diploid cultures of *Saccharomyces cerevisiae* exposed to varying doses of X-radiation. A preliminary plot indicated that doses equal to or greater than 40,000 roentgens would satisfy approximation (2). Only these points have been used in the calculations.

Although at first glance this technique seems satisfactory and appealing because of the simplicity of the calculations, closer scrutiny reveals that n and k are actually very poorly estimated. Since the organism is a diploid we know that $n = 2$. But if this information were not known, the 95% confidence interval for $\hat{n}$, which is (1.8, 3.7) would tell us only that it is a diploid or a triploid, possibly even a tetraploid. Furthermore, as the number of sensitive units per organism increases, the confidence intervals computed from this method become progressively worse. Actually the poor quality of the estimation is only a reflection of the well known fact that confidence intervals for a regression line are extremely wide outside the range of the observed data. In other words the gain in simplicity achieved by virtue of approximation (2) is more than offset by an increase in the errors of estimate. The same argument without modification cannot be applied to the estimation of k by this method, but as it turns out in this example, k is also poorly estimated. Clearly, some approach which permits the use of all points on the curve must be used.

As an alternative to (3) we may write

$$u_i = \log (1 - S_i) = n \log (1 - e^{-kx_i}) + \epsilon_i \tag{4}$$

which involves no approximations. The estimation procedure requires the minimization of the quantity

$$V = \sum_{i=1}^{p} [u_i - n \log (1 - e^{-kx_i})]^2$$

for variations in n and k. The equations for $\hat{n}$ and $\hat{k}$ so obtained are non-linear and in general must be handled by iterative methods. However, for most experiments, a short-cut method is available which ordinarily will give reliable results. Stepwise the method may be described as follows:

1. Select a trial value of k, say k_0, which may be obtained from a plot of the higher dose points.

2. Find the value of n, say n_0, which minimizes V for $k = k_0$. It is

$$n_0 = \sum uv / \sum v^2$$

where

$$v = \log (1 - e^{-k_0 x}).$$

3. Compute $V_{\min} = \sum u^2 - n_0 \sum uv$.

4. Repeat the process for two or three more trial values of k, chosen so as to bracket the absolute minimum of $V_{\min}$.

5. Plot each V_{min} against its corresponding trial value of k and interpolate to obtain the value of k which makes V an absolute minimum.

6. This final value of k and the corresponding n computed from the formula in step 2 are the desired estimates $\hat{k}$ and $\hat{n}$.

In selecting an initial trial value of k, it should be remembered that the estimate of k obtained from the semi-logarithmic plot of the higher dose points will usually be too large. In Pomper's data the estimate was 0.61, so the initial trial value chosen was 0.57. The computations

TABLE 3

FITTING OF A MULTI-HIT CURVE, EQ. (4)

[Data from Pomper (1952)]

Proportion killed	log $(1 - S)$			log $(1 - e^{-.57x})$
$1 - S$	u	$.57x$	$1 - e^{-.57x}$	v
.14	−1.9661	.57	.4345	−.8336
.36	−1.0217	1.14	.6802	−.3854
.79	−0.2357	2.28	.8977	−.1079
.928	−0.0747	3.42	.9673	−.0333
.980	−0.0202	4.56	.9895	−.0105
.9944	−0.0056	5.70	.9966	−.0034
.9980	−0.0020	6.84	.9989	−.0011
.99955	−0.0004	7.98	.9997	−.0003

$$\sum u^2 = 4.971023, \qquad \sum v^2 = .856210 \qquad \sum uv = 2.060758$$

$$n_0 = 2.060758/.856210 = 2.406837$$

$$V_{min} = 4.971023 - 2.406837(2.060758) = .01111$$

RESULTS OF SUBSEQUENT TRIALS

k	V_{min}
.55	.00931
.50	.00938

for steps 2 and 3 are shown in Table 3. Note that the quantity, $\sum u^2$, has to be computed only once, since it does not involve k. Complete calculations are given only for $k = 0.57$. The results for two more trial values are given at the bottom of the table. The points are plotted in

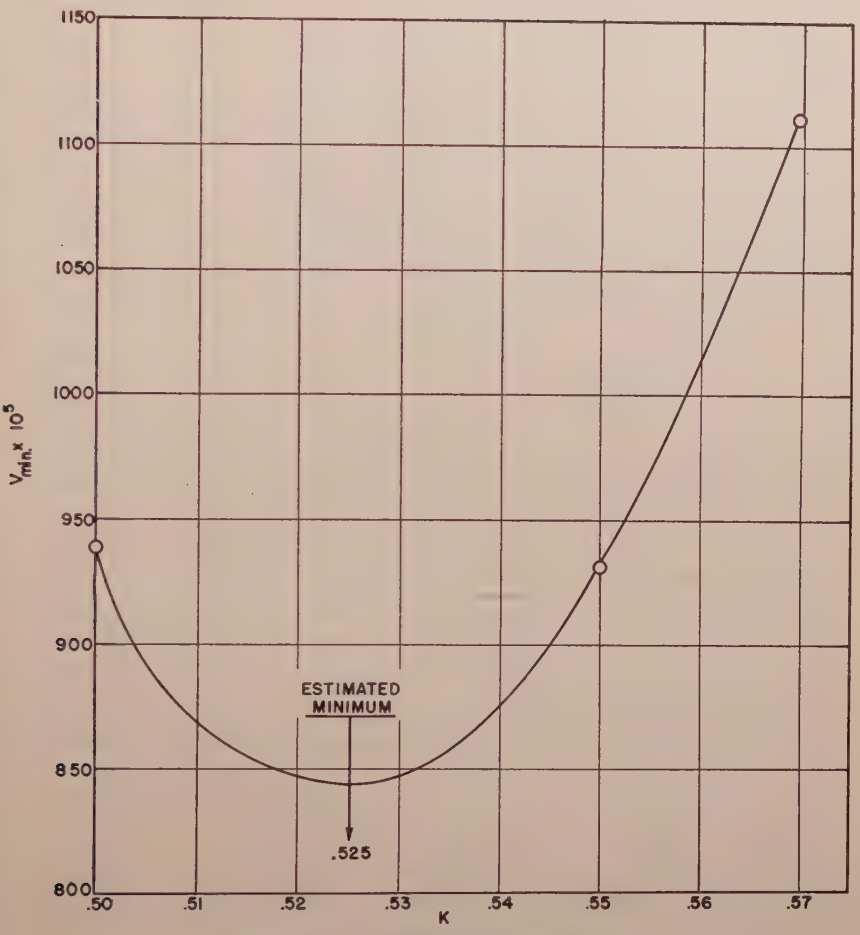

FIGURE 1 GRAPHICAL ESTIMATION OF K

Figure 1, and the curve drawn free hand through the points indicates an absolute minimum at about $k = 0.525$. Using this as the final $\hat{k}$ we find $\sum v^2 = 1.0065$, $\sum uv = 2.2349$ and $\hat{n} = 2.2205$. These estimates were found to satisfy the true estimation equations to three significant figures.

The computation of standard errors for estimates $\hat{n}$ and $\hat{k}$ obtained from model (4) is complicated by the fact that the estimates are not linear functions of the u_i . In this case the only practical method of obtaining error estimates is to compute the asymptotic variance-covariance matrix. The formulas, derivation of which is outlined in section 4, are

$$s_{\hat{n}}^2 = s^2[B/(AB - C^2)]$$
$$s_{\hat{k}}^2 = s^2[A/(AB - C^2)]$$

where

$$A = \sum v^2$$

$$B = \hat{n}^2 \sum \left[\frac{xe^{-\hat{k}x}}{(1 - e^{-\hat{k}x})} \right]^2$$

$$C = \hat{n} \sum \left[\frac{vxe^{-\hat{k}x}}{(1 - e^{-\hat{k}x})} \right]$$

$$s^2 = \frac{1}{p - 2} [\sum u^2 - \hat{n} \sum uv].$$

TABLE 4

CALCULATIONS FOR STANDARD ERRORS OF $\hat{n}$ AND $\hat{k}$, EQ. (4)

(1)	(2)	(3)	(4) = (3) ÷ [1 − (2)]	(5)
x	$e^{-\hat{k}x}$	$xe^{-\hat{k}x}$	$\frac{xe^{-\hat{k}x}}{(1 - e^{-\hat{k}x})}$	v
1	.5916	.5916	1.4483	−.8954
2	.3499	.6999	1.0766	−.4307
4	.1225	.4898	.5582	−.1306
6	.0428	.2571	.2686	−.0438
8	.0150	.1200	.1218	−.0151
10	.0052	.0525	.0528	−.0053
12	.0018	.0221	.0221	−.0018
14	.0006	.0090	.0090	−.0006

$$A = \sum (5)^2 = 1.0065$$

$$B = \hat{n}^2 \sum (4)^2 = (2.2205)^2(3.6586)$$

$$= 18.0390$$

$$C = \hat{n}^2 \sum (4)(5) = (2.2205)(-1.8473)$$

$$= -4.1020$$

$$s^2 = \tfrac{1}{6}[4.971023 - 2.220471(2.234919)]$$

$$= .001408$$

The calculation of A, B and C is carried out in Table 4. The columns are set up in an orderly fashion and B and C can be obtained from sums of squares or cross-products of the appropriate columns. Other tabular arrangements may be found more suitable for some computers. Finally the standard errors are found to be $s_{\hat{n}} = .14$, $s_{\hat{k}} = .033$. Although the number of degrees of freedom to be associated with s^2 is not well defined, it is suggested that for purposes of computing confidence intervals, $(p - 2)$ be used in conjunction with Student's "t" in the usual manner. When this is done the approximate 95% confidence intervals for $\hat{n}$ and $\hat{k}$ are (1.9, 2.6) and (.44, .60), respectively.

The computational labor in fitting model (4) is greater than in fitting model (3), but it is not excessive. In any particular case, the additional work needed to obtain more accurate estimates must be balanced against the consequences of possibly incorrect or misleading conclusions which might result from the simpler analysis. In the example just discussed, the simpler analysis might have led to the erroneous conclusion that the original culture was a triploid and would have overestimated k. Whether such errors can be tolerated will depend on the individual experimenter and the significance of a particular experiment.

More general models have been discussed by Atwood and Norman (1949). The population may consist of organisms each with a different number of sensitive units. In *Neurospora crassa*, for example, populations of macroconidia are found in which cells have different numbers of nuclei. In this case the method of fitting just described estimates the average number of nuclei per cell in the population. A further generality is introduced if the sensitivities of the n units in an organism are not the same, i.e., if each unit has a different k. The fitting of curves based on this model would necessitate the use of an iterative solution of estimation equations. Because of the length and complexity of calculations required for such methods, their use is limited.

4. *Some comments on the theory.*

The statistical methods employed in the previous sections depend first of all on the assumption that the ϵ_i are independent normally distributed random variables with zero expectations and a variance which is the same for all i. The truth is that they are independent and do have zero expectations, but they are not normally distributed and do not have uniform variances. In most cases the survival proportions are obtained from the ratio of two bacterial counts which suggests that the ϵ_i might behave somewhat like the ratio of two Poisson variates. Conceivably one could find a transformation which would render the variances nearly uniform and perhaps induce approximate normality, but it would likely

have two disadvantages. It would probably do poorly at very high or very low survivals, the latter having occurred in the example just discussed, and it would make the introduction of a specific model relating survival and dose even more difficult than it already is. The problem of heterogeneous variances is often handled quite adequately by the experimenter simply by increasing the number of plates or loops at doses which are expected to have ϵ_i with high variances. One never achieves perfect homoscedasticity by such procedures but they usually suffice for most practical purposes. As for the lack of normality, the logarithm of the ratio of two Poisson variates in which the denominator is determined much more accurately than the numerator might be expected to have a moderately symmetrical distribution and therefore not depart so far from normality as to cause any serious difficulties.

The standard errors of $\hat{n}$ and $\hat{k}$ given in the last section are obtained from the asymptotic variance-covariance matrix which is derived in the usual manner. If the ϵ_i satisfy the assumptions in the previous paragraph, the probability of the sample is

$$P(S) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^p \exp\left\{-\frac{1}{2}\sigma^2 \sum_{i=1}^{p} [u_i - n \log (1 - e^{-kx_i})]^2\right\}$$

and the likelihood is $L = \log P(S)$. Then the asymptotic variance-covariance matrix is the inverse of the matrix

$$-E\begin{bmatrix} \frac{\partial^2 L}{\partial n^2} & \frac{\partial^2 L}{\partial n\, \partial k} \\ \frac{\partial^2 L}{\partial n\, \partial k} & \frac{\partial^2 L}{\partial k^2} \end{bmatrix} = \frac{1}{\sigma^2}\begin{bmatrix} A & C \\ C & B \end{bmatrix}$$

where A, B and C are defined as in section 3. In practice, σ^2, n and k must be replaced by their estimates, s^2, $\hat{n}$ and $\hat{k}$. This of course means that the standard error estimates are only approximate, and a further ambiguity is introduced when an attempt is made to estimate confidence intervals for n and k. One reasonably safe procedure is to associate $(p - 2)$ degrees of freedom with s^2 and to compute confidenec intervals using Student's "t". It is difficult to determine the accuracy of such methods. A rough check, however, seems to indicate that they are conservative. If k is assumed to be known and equal to 0.525, the variance of $\hat{n}$, which is then properly estimated, is 0.001396. The variance of $\hat{n}$ computed from the asymptotic variance-covariance matrix is over ten times as large as this, a fact which would seem to support the contention that the approximation is on the conservative side.

5. *Summary.*

The fitting of theoretical survival curves to data obtained from radiation experiments with microorganisms is discussed from a statistical point of view. Formulas are given for obtaining estimates of parameters and standard errors. The methods are illustrated with two sets of experimental data.

The author gratefully acknowledges many helpful discussions with Dr. K. C. Atwood and is indebted to Dr. S. Pomper for permission to use his hitherto unpublished data.

REFERENCES

Atwood, K. C. and A. Norman. On the interpretation of multi-hit survival curves. *Proc. N. Y. Acad. Sci., 35(12):* 696–709. 1949.

Lea, D. E., R. B. Haines and C. A. Coulson. The mechanism of the bactericidal action of radioactive radiations. *Proc. Roy. Soc. Lond.*, B, *120:* 47–76. 1936.

Pomper, S. Unpublished data. 1952.

POPULATION GROWTH OF THE SEXES

LEO A. GOODMAN[1]

University of Chicago

1. INTRODUCTION

Many authors (see, e.g., references 1–19) have noted that the sex distribution at birth is not equal (more males than females are born) and that the ability of the sexes to withstand the forces of mortality is not the same (the death rates at most ages are higher for males than for females). This phenomenon is not confined to man alone, but it also occurs, though not universally, among a number of other species. The literature contains many discussions of the importance and implications of these two facts.

At a recent meeting of the Royal Statistical Society several speakers pointed out that the possibility of variations in the relative numbers of the two sexes has been too long neglected in population mathematics. D. G. Kendall [22] has discussed some of the characteristic difficulties of this problem and suggested some approximations which we shall extend. Some work on this problem, from a deterministic standpoint, has been carried out (see, e.g., references 1, 13–19). Kendall has mentioned that the problem of expressing the modes of population growth for the two sexes in stochastic form would be a very difficult one. References 20–26 deal with this problem when one sex is considered and their bibliographies give references to much of what has been written in the field.

J. Yerushalmy, in a very interesting paper [1], describes the age-sex composition of the population resulting from natality and mortality conditions. If we are interested in the ultimate stationary population we may determine its over-all sex ratio from Yerushalmy's analysis. In this paper we shall consider the problem of determining the over-all sex ratio for populations which need not be stationary and also the problem of studying the population growth of each sex.

We shall consider various mathematical models of population growth. In order to make possible an analytic treatment of the subject, these models necessarily will be oversimplifications of reality.

---

[1]This paper was prepared in connection with research supported by the Office of Naval Research.

2. DETERMINISTIC MODELS

Extending Kendall's [22, p. 247] approach we have the following model: If $M(t)$ and $F(t)$ are the number of males and females respectively at time t, then these quantities satisfy the differential equations

$$\frac{dM}{dt} = -aM + \Lambda(M, F)$$

$$\frac{dF}{dt} = -bF + \Lambda'(M, F)$$

where a is the intrinsic male death rate per male per unit of time, b is the intrinsic female death rate per female per unit of time, $\Lambda(M,F)$ and $\Lambda'(M,F)$ are functions of M and F representing the contributions from the male birth rate and female birth rate, respectively. (Kendall dealt with the case where $a = b$ and $\Lambda = \Lambda'$.)

Consider the case where $\Lambda(M,F) = uF$ and $\Lambda'(M,F) = vF$; i.e., where the birth rates depend on the female population size F (females are marriage dominant, see [17], [18]). We then have

$$\frac{dM}{dt} = -aM + uF$$

$$\frac{dF}{dt} = -bF + vF = (v - b)F.$$

The solution of these equations is

$$M(t) = \frac{uA}{(v - b + a)} e^{(v-b)t} + Be^{-at}$$

$$F(t) = Ae^{(v-b)t},$$

where A and B are determined by the population composition at $t = 0$. We have that the sex ratio is

$$\frac{M(t)}{F(t)} = S(t) = \left\{1 + \frac{B(v - b + a)}{uA} e^{-(v-b+a)t}\right\} \Big/ \left[\frac{v - b + a}{u}\right],$$

and the ultimate sex ratio is $S = S(\infty) = u/[v + a - b]$, when $v + a - b > 0$. We note that M and F tend jointly to infinity when $v - b > 0$, and to zero when $v - b < 0$. When the female death rate equals the female birth rate, $v - b = 0$, M and F tend to nonzero limits and the sex ratio approaches $S = u/a$ the ratio between the male birth rate and the male death rate.

Now consider the case where $\Lambda(M,F) = uM$ and $\Lambda'(M,F) = vM$; i.e., where the birth rates depend on the male population (males are

marriage dominant). By applying the methods of the preceding case we may determine $M(t)$, $F(t)$, $S(t)$, and we find

$$S = \frac{u - a + b}{v},$$

when $u - a + b > 0$. Also, M and F behave qualitatively in the same way as before. When $u - a = 0$, M and F tend to nonzero limits, and $S = b/v$.

Let us consider the case where

$$\Lambda(M, F) = u\left(\frac{M + F}{2}\right)$$

and

$$\Lambda'(M, F) = \frac{v(M + F)}{2};$$

i.e., where the contributions from the birth rates depend on the total population size $M + F$ (neither males nor females are marriage dominant). We then have

$$\frac{dM}{dt} = -aM + \frac{u}{2}(M + F) = \left(-a + \frac{u}{2}\right)M + \frac{u}{2}F = fM + cF$$

$$\frac{dF}{dt} = -bF + \frac{v}{2}(M + F) = \left(-b + \frac{v}{2}\right)F + \frac{v}{2}M = gF + kM,$$

where

$$f = \frac{u}{2} - a; \qquad g = \frac{v}{2} - b, \qquad c = \frac{u}{2}, \qquad k = \frac{v}{2}.$$

The solution of these equations is

$$M(t) = Ae^{[f+g+h]t/2} + \left\{\frac{f - g - h}{2k}\right\}Be^{[f+g-h]t/2},$$

$$F(t) = \left\{\frac{-f + g + h}{2c}\right\}Ae^{[f+g+h]t/2} + Be^{[f+g-h]t/2},$$

where $h = +\sqrt{(f - g)^2 + 4ck}$, and A, B are determined by the population composition at $t = 0$. Hence the sex ratio is

$$S(t) = \frac{M(t)}{F(t)} = \frac{1 + \left\{\frac{f - g - h}{2k}\right\}\frac{B}{A}e^{-ht}}{\left\{\frac{-f + g + h}{2c}\right\} + \frac{B}{A}e^{-ht}}$$

and the ultimate sex ratio is

$$S = S(\infty) = 2c/[g - f + h]$$

$$= u \Big/ \left[\frac{v-u}{2} + a - b + \sqrt{\left(\frac{v-u}{2} + a - b \right)^2 + uv} \right].$$

We could have determined that S was one of two values directly by noting that

$$dS/dt = d\left(\frac{M}{F}\right) \Big/ dt = \left[\frac{F\,dM - M\,dF}{F^2} \right] \Big/ dt$$

$$= \left[\frac{fMF + cF^2 - gFM - kM^2}{F^2} \right]$$

$$= (f - g)S + c - kS^2.$$

The solution of this equation for $dS/dt = 0$ is

$$\frac{g - f \pm \sqrt{(g-f)^2 + 4kc}}{-2k} = \frac{g - f \pm h}{-2k}$$

$$\left(\text{we have } \frac{g - f - h}{-2k} = \frac{2c}{g - f + h} = S \right).$$

We note that M and F tend jointly to infinity when $f + g + h > 0$, and to zero when $f + g + h < 0$. When $f + g + h = 0$, M and F tend to nonzero limits and $S = u/-2f = u/(2a - u) = (2b - v)/v$. The critical relation between the constants is

$$f + g = -h$$

$$(f + g)^2 = (f - g)^2 + 4kc$$

$$fg = kc,$$

or

$$(2a - u)(2b - v) = uv$$

$$a(v - b) + b(u - a) = 0$$

which is a generalization of Kendall's critical relation [22, p. 248].

Let us now consider the case where $\Lambda(M,F) = u\sqrt{MF}$ and $\Lambda'(M,F) = v\sqrt{MF}$, i.e., where the contributions from the birth rates depend on the geometric mean of M and F. We then have

$$\frac{dM}{dt} = -aM + u\sqrt{MF}$$

$$\frac{dF}{dt} = -bF + v\sqrt{MF}.$$

Writing $\sqrt{M} = X$, $\sqrt{F} = Y$,

$$2X\frac{dX}{dt} = -aX^2 + uXY$$

$$\frac{dX}{dt} = fX + cY$$

and

$$2Y\frac{dY}{dt} = -bY^2 + vXY$$

$$\frac{dY}{dt} = gY + kX,$$

where $-a/2 = f$, $-b/2 = g$, $u/2 = c$, $v/2 = k$. Hence using the results for the preceding model, we can compute $X(t)$, $Y(t)$, $X^2(t) = M(t)$, $Y^2(t) = F(t)$, and $S(t)$ directly. We find that

$$S = S(\infty) = \{2c/[g - f + h]\}^2$$

$$= \left\{u \Big/ \left[\frac{a-b}{2} + \sqrt{\left(\frac{a-b}{2}\right)^2 + uv}\right]\right\}^2.$$

Also M and F behave qualitatively in the same way as before, the critical relation between the constants being

$$fg = kc, \qquad \text{or} \qquad ab = uv,$$

in which case

$$S = [u/-2f]^2 = [u/a]^2 = [b/v]^2.$$

It is interesting to note that in all the models we have considered heretofore, when the population size tends to a nonzero limit, the sex ratio is what one would expect of a stationary situation; i.e., $S = bu/av$, which equals u/a when $v - b = 0$, and b/v when $u - a = 0$, and $(b/v)^2$ when $ab = uv$ (in the preceding model).

Consider now a model which discriminates between married and unmarried persons. Let M, F, and N denote at time t the number of unmarried males, unmarried females, and married couples, respectively. Then a natural generalization of the preceding models is governed by the equations

$$\frac{dM}{dt} = -aM + uN + nN - K(M,F)$$

$$\frac{dF}{dt} = -bF + vN + mN - K(M,F)$$

$$\frac{dN}{dt} = -(m+n)N + K(M,F)$$

where a is the unmarried male death rate per unmarried male per unit of time, b is the unmarried female death rate per unmarried female per unit of time, u is the male birth rate per married couple per unit of time, v is the female birth rate per married couple per unit of time, m is the married male death rate per married couple per unit of time, n is the married female death rate per married couple per unit of time, and $K(M,F)$ is the marriage rate per unit of time (the age distribution is, of course, being neglected throughout).

In the case where, $K(M,F) = cF$ (female marriage dominance), the equations reduce to

$$\frac{dM}{dt} = -aM + dN - cF$$

$$\frac{dF}{dt} = gF + kN$$

$$\frac{dN}{dt} = fN + eF,$$

where $d = u + n$, $g = -b - c$, $k = v + m$, $f = -(m + n)$. The solution of these equations is

$$N(t) = Ae^{[f+g+h]t/2} + \left\{\frac{f - g - h}{2k}\right\}Be^{[f+g-h]t/2}$$

$$F(t) = \left\{\frac{-f + g + h}{2c}\right\}Ae^{[f+g+h]t/2} + Be^{[f+g-h]t/2}$$

$$M(t) = A[2d + f - g - h]e^{[f+g+h]t/2}/[f + g + h + 2a]$$
$$+ B[d(f - g - h)/k - 2c]e^{[f+g-h]t/2}/[f + g - h + 2a]$$
$$+ De^{-at}$$

where $h = +\sqrt{(f - g)^2 + 4ck}$, and A, B, D are determined by the population composition at $t = 0$.

Hence the sex ratio of unmarried people

$$S'(t) = \frac{M(t)}{F(t)}$$

and the overall sex ratio

$$S(t) = \frac{M(t) + N(t)}{F(t) + N(t)}$$

may be evaluated. We find that

$$S'(\infty) = [2d + f - g - h]2c/[f + g + h + 2a][g - f + h]$$

and

$$S(\infty) = \left[\left(\frac{2d + f - g - h}{f + g + h + 2a}\right) + 1\right] \Big/ \left[\frac{g - f + h}{2c} + 1\right].$$

We also note that M and F behave qualitatively in the same way as before, the critical relation between the constants being

$$fg = kc, \qquad \text{or} \qquad (m + n)(b + c) = (v + m)c,$$

in which case,

$$\begin{aligned} S'(\infty) &= -(d + f)c/af = -(d + f)g/ak \\ &= (u - m)(b + c)/a(v + m), \end{aligned}$$

and

$$S(\infty) = (d + f + a)c/a(c - f) = \frac{(d + f + a)(m + n)(b + c)}{(v + m)a(c - f)}$$

$$= (u - m + a)(b + c)(m + n)/a(v + m)(c + m + n).$$

We see that the sex ratio differs from what one obtains in a simple stationary situation.

A similar analysis could be carried out for models which discriminate between married and unmarried persons and where the males are marriage dominant ($K(M,F) = cM$) or where neither sex is marriage dominant (e.g., $K(M,F) = c\sqrt{MF}$ or $K(M,F) = c(M + F)/2$).

3. SIMPLE STOCHASTIC MODELS

Let us consider a slight generalization of a stochastic model described by Moran [25]. If $B(t)$ is the total number of males, and $G(t)$ the total number of females born since time $t = 0$, and if $B(t) = c$, and $G(t) = g$ at time t, the probabilities at time $t + dt$ are given by

$$\begin{aligned} Pr\{B(t + dt) = c + 1\} &= u\,dt + o(dt) \\ Pr\{B(t + dt) = c\} &= 1 - u\,dt + o(dt) \\ Pr\{G(t + dt) = g + 1\} &= v\,dt + o(dt) \\ Pr\{G(t + dt) = g\} &= 1 - v\,dt + o(dt), \end{aligned}$$

where B and G are independent. The distribution of length of life may be of any general form and we suppose that the average length of life for the males is $L_M = 1/a$, and the average length of life for the females is $L_F = 1/b$. Then it is not too difficult to see that if $M(t)$ is the expected number of males alive at time t, and $F(t)$ is the expected number of females alive at time t, then the sex ratio will ultimately be

$$S(\infty) = \lim_{t\to\infty} \frac{M(t)}{F(t)} = \frac{uL_M}{vL_F} = bu/av.$$

This result is the same as that of the corresponding deterministic model

$$\frac{dM}{dt} = -aM + u$$

$$\frac{dF}{dt} = -bF + v.$$

Now consider the following process: In order that the total population size remain constant, when a person dies there is a "replacement" made (a birth takes place) which will be either a male or female with chances $u/(u + v)$ and $v/(u + v)$, respectively. The chance that a given male will die in a unit of time is a, and b is the chance that a given female will die. We might now consider a given male and study its life span and also the life span of its "replacement," and the life span of the "replacement's replacement," etc. The probabilities at the tth unit of time are given by

$$Pr_t\{M\} = Pr_{t-1}\{M\}(1 - a) + Pr_{t-1}\{M\}a \frac{u}{u+v} + Pr_{t-1}\{F\}b \frac{u}{u+v},$$

where $Pr_t\{M\}$ is the probability that the "replacement" at the tth unit of time for a given male is a male, $Pr_t\{F\}$ is the probability that the "replacement" at the tth unit of time for a given male is a female, $Pr_t\{M\} + Pr_t\{F\} = 1$. The value of $Pr_t\{M\}$ may be computed explicitly as a function of a, b, u, v, and t. As t approaches infinity, we find that

$$Pr_\infty\{M\}a \frac{v}{(u+v)} = Pr_\infty\{F\}b \frac{u}{(u+v)},$$

and, hence,

$$\frac{Pr_\infty\{M\}}{Pr_\infty\{F\}} = \frac{bu}{av}.$$

We consider only the case where $Pr_0\{M\} = 1$, since the more general

case may be analyzed using only the special case. Since the total population size is a constant N, the expected number of males alive in the tth unit of time is $N\ Pr_t\{M\}$, and the expected number of females alive in the tth unit of time is $N\ Pr_t\{F\}$. Therefore, the overall sex ratio for the population will ultimately be

$$S(\infty) = bu/av.$$

The model we have just considered is a special case in the general theory of renewal. An excellent exposition of this topic appears in [26]. It may be shown that more general assumptions concerning the distribution of length of life still lead to the same results when t approaches infinity.

The preceding simple stochastic models have been presented more as an illustration of the problem than as a solution to it. We shall now consider more complicated stochastic models.

4. STOCHASTIC MODELS OF POPULATION GROWTH

We shall now consider a stochastic process which is a probabilistic analogue of the deterministic model describing a population where the females are marriage dominant; i.e., where the birth rates depend on the female population size.[1] Although the analysis which follows pertains to populations where the females are marriage dominant, the results obtained may be used in an obvious manner to analyze populations where the males are marriage dominant.

The population under consideration behaves in accordance with the following rules:

(a) the subpopulations generated by two co-existing individuals develop in complete independence of one another;

(b) a female alive at time t has a chance $u\ dt + o(dt)$ of giving birth to a male and a chance $v\ dt + o(dt)$ of giving birth to a female during the following time interval of length dt;

(c) a female at time t has a chance $b\ dt + o(dt)$ of dying and a male alive at time t has a chance $a\ dt + o(dt)$ of dying in the following time interval of length dt.

Let the random variables $M(t)$ and $N(t)$ denote the number of males and females, respectively, alive at time t. We see that the female population follows the rules of a simple birth-and-death process (cf. [22], p. 236), the mean female population size being

$$\overline{N}(t) = e^{(v-b)t}$$

[1]The author wishes to express his indebtedness to David G. Kendall of Magdalen College, Oxford and Princeton University, and Charles Stein of the University of Chicago for their assistance in the analysis of this stochastic process.

while

$$\text{Var}\{N(t)\} = \frac{v+b}{v-b}e^{(v-b)t}\{e^{(v-b)t} - 1\},$$

when $N(0) = 1$. Since the subpopulations generated by two coexisting individuals develop in complete independence of one another, we need only multiply $\overline{N}(t)$ and $\text{Var}\{N(t)\}$ by $N(0)$ in the more general case where there are $N(0) \geq 1$ females alive at the initial time $t = 0$. There will be "almost certain" extinction unless the mean female population size is increasing ($v > b$). When $v = b$ and the mean female population size is constant, we still find that there will be "almost certain" extinction.

The behavior of the male population is somewhat more difficult to analyze since male births depend on the female population. The method of analysis will be described in the Appendix. We find that the mean male population size is

$$\overline{M}(t) = \frac{u}{(v-b+a)}[e^{(v-b)t} - e^{-at}]$$

while,

$$\text{Var}\{M(t)\} = \frac{(v+b)u^2}{(v-b+a)^2}\left\{\frac{e^{2(v-b)t}}{(v-b)} - \frac{e^{-2at}}{(v-b+2a)} + \frac{2e^{(v-b-a)t}}{a}\right\}$$
$$+ ue^{(v-b)t}\left\{\frac{1}{(v-b+a)} - \frac{2u(v+b)}{a(v-b)(v-b+2a)}\right\} - \frac{ue^{-at}}{(v-b+a)},$$

when $M(0) = 0$ and $N(0) = 1$. In the more general case where $N(0) \geq 1$ females are alive at the initial time $t = 0$, the formulas $\overline{M}(t)$ and $\text{Var}\{M(t)\}$ are multiplied by $N(0)$, since the subpopulations generate independently,

In the more general case where there are $M(0) \geq 0$ males at the initial time $t = 0$, these $M(0)$ males follow the rules of a simple death process; i.e., the chance that $M_0(t)$ from among the original $M(0)$ males will be alive at time t is

$$C_{M_0(t)}^{M(0)}p(t)^{M_0(t)}q(t)^{M(0)-M_0(t)},$$

where $p(t) = 1 - q(t) = e^{-at}$. Hence,

$$\overline{M}_0(t) = M(0)e^{-at}$$

and

$$\text{Var}\{M_0(t)\} = M(0)e^{-at}[1 - e^{-at}].$$

We then add $\overline{M}_0(t)$ and $\mathrm{Var}\{M_0(t)\}$ to $\overline{M}(t)$ and $\mathrm{Var}\{M(t)\}$, respectively, to obtain the mean and variance of the total number of males when there are $M(0)$ males and $N(0)$ females at time $t = 0$. These additions have, of course, no influence on the form of the solution when $t \to \infty$.

We find that the covariance between $M(t)$ and $N(t)$ is

$$\mathrm{Cov}\,(t) = E\{[M(t) - \overline{M}(t)][N(t) - \overline{N}(t)]\}$$

$$= \frac{u(v+b)}{(v-b+a)}\left\{\frac{e^{2(v-b)t}}{(v-b)} + \frac{e^{(v-b-a)t}}{a}\right\} - \frac{u}{a}\frac{(v+b)}{(v-b)}e^{(v-b)t},$$

when $M(0) = 0$ and $N(0) = 1$. In the more general case where $N(0) \geq 1$, $M(0) \geq 0$, then the covariance is multiplied by $N(0)$.

In the special case where the constants are equal ($a = b = u = v = \lambda$), we have

$$\overline{N}(t) = N(0), \qquad \overline{M}(t) = N(0)[1 - e^{-\lambda t}] + M(0)e^{-\lambda t},$$

$$\mathrm{Var}\,\{N(t)\} = N(0)2\lambda t,$$

$$\mathrm{Var}\,\{M(t)\} = N(0)[2\lambda t - 2 + 3e^{-\lambda t} - e^{-2\lambda t}] + M(0)e^{-\lambda t}[1 - e^{-\lambda t}],$$

$$\mathrm{Cov}\,(t) = N(0)[2\lambda t - 2 + 2e^{-\lambda t}],$$

and the correlation coefficient $\rho(t) = \mathrm{Cov}(t)/\sqrt{\mathrm{Var}\,\{M(t)\}\ \mathrm{Var}\,\{N(t)\}}$ between the number of males and the number of females in the population is seen to approach one as t becomes large.

The methods which we have used to analyze the stochastic process representing a marriage dominant (female or male) population, can be extended in order to analyze a stochastic process where the contributions from the birth rates depend on the total population size (neither male nor female is marriage dominant). That is, the population under consideration behaves in accordance with the following rules:

(a) the subpopulation generated by two coexisting individuals develops in complete independence of one another;

(b) an individual alive at time t has a chance $u\,dt/2 + o(dt)$ of reproducing a male and a chance $v\,dt/2 + o(dt)$ of reproducing a female during the following time interval of length dt;

(c) a female alive at time t has a chance $b\,dt + o(dt)$ of dying and a male alive at time t has a chance $a\,dt + o(dt)$ of dying in the following time interval of length dt.

Using the second method described in the Appendix we may obtain the moments and cross moments of $M(t)$ and $N(t)$, the number of males and females, respectively, alive at time t. The means, variances, and the

covariance of $M(t)$ and $N(t)$ satisfy a system of five differential equations with constant coefficients which may be solved to determine these moments explicitly.

Let us consider the special case where the constants are all equal ($a = b = u = v = \lambda$). Then the probability $P_{m,n}(t)$ that at time t the population contains m males and n females may be determined explicitly when the initial conditions are $P_{0,1}(0) = P_{1,0}(0) = 1/2$. We find that

$$P_{m,n}(t) = C_m^{m+n}\left(\frac{1}{1+\lambda t}\right)^2\left(\frac{\lambda t}{1+\lambda t}\right)^{m+n-1}\left(\frac{1}{2}\right)^{m+n}$$

for $m + n \geq 1$, and

$P_{0,0}(t) = \lambda t/(1 + \lambda t)$. Whence we see that there will be "almost certain" extinction. The probability $P_m(t)$ that at time t the population contains m males, and the probability $P'_n(t)$ that at time t the population contains n females may also be determined explicitly. We have

$$P_m(t) = P'_m(t) = \left(\frac{\lambda t}{2+\lambda t}\right)^{m-1} 2/(2 + \lambda t)^2,$$

for $m \geq 1$ and

$$P_0(t) = P'_0(t) = (1 + \lambda t)/(2 + \lambda t).$$

We find that $\overline{M}(t) = \overline{N}(t) = 1/2$, $\mathrm{Var}\{M(t)\} = \mathrm{Var}\{N(t)\} = (2\lambda t + 1)/4$, $\mathrm{Cov}(t) = (2\lambda t - 1)/4$, and the correlation coefficient $\rho(t) = (2\lambda t - 1)/(2\lambda t + 1)$.

5. APPENDIX

In studying the stochastic process representing a population where the females are marriage dominant, the means, variances, and covariances of the male and female population sizes were given. The following method which was used to obtain the first and second moments could also be used to obtain higher moments:

All moments for the female population size may be obtained directly from its distribution which is a geometric series with a modified zero term (cf. [22] p. 237) or from its moment generating function. All moments for the male population size may be obtained from its moment generating function $\varphi(z,t) = E\{z^{M(t)}\}$. We find that $\varphi(z,t)$ satisfies the differential equation

$$\frac{\partial \varphi}{\partial t} = (b - v\varphi)(1 - \varphi) + u\varphi(z - 1)e^{-at} \tag{1}$$

and that $\varphi(z,0) = z$, when $M(0) = 0$ and $N(0) = 1$. By differentiating (1) once and setting $z = 1$, we obtain a linear differential equation for the mean $\overline{M}(t)$. By differentiating (1) several times and setting $z = 1$, we obtain linear differential equations for the higher factorial moments of $M(t)$.

The method just described gives the moments of $M(t)$ but does not give the cross moments of $M(t)$ and $N(t)$; e.g., the covariance. Another method might be used to obtain the moments and cross moments of $M(t)$ and $N(t)$ which serves as an independent check on the preceding method. The two methods might be used simultaneously to simplify computation. Let $P_{m,n}(t)$ be the probability that at time t the population contains m males and n females. Then the infinite system of functions $P_{m,n}(t)$, $m = 0, 1, 2, \cdots$, $n = 0, 1, 2, \cdots$, satisfy a basic system of ordinary differential equations (cf. [23], Chap. 17). By multiplying each differential equation by an appropriate factor and then adding the entire system of differential equations, we obtain ordinary differential equations for the moments; e.g., if we multiply the differential equation containing the term $[\partial P_{m,n}(t)]/\partial t$ by m and then add the entire system of differential equations, we will obtain an equation containing the term

$$\sum_{m,n=0}^{\infty} m \frac{\partial P_{m,n}(t)}{\partial t}$$

which is in fact $[\partial \overline{M}(t)]/\partial t$. We then have a differential equation in $\overline{M}(t)$ which may be solved directly.

REFERENCES

1. J. Yerushalmy, "The age-sex composition of the population resulting from natality and mortality conditions," *The Milbank Memorial Fund Quarterly*, Vol. XXI (1943), No. 1, pp. 37–63.
2. Sanford Winston, "The influence of social factors upon the sex ratio at birth," *The American Journal of Sociology*, Vol. 37 (1931), pp. 1–21.
3. Sanford Winston, "Birth control and the sex ratio at birth," *The American Journal of Sociology*, Vol. 38 (1932), pp. 225–231.
4. Christopher Tietze, "A note on the sex ratio of abortions," *Human Biology*, Vol. 20 (1948), pp. 156–160.
5. R. J. Meyers, "Effect of the war on the sex ratio at birth," *American Sociological Review*, Vol. 12 (1947), No. 1, pp. 40–43.
6. Rachel M. Jenss, "An inquiry into methods of studying the sex ratio at birth for the United States during war and postwar years," *Human Biology*, Vol. 15 (1943), No. 3, pp. 255–266.
7. A. S. Parkes, "The physiological factors governing the proportions of the sexes in man," *The Eugenics Review*, Vol. XVII (1926), pp. 275–291.
8. A. S. Parkes, "The mammalian sex ratio," *Biological Review*, Vol. 1 (1926), p. 1.
9. H. Stranskov, "On the variance of human live birth sex ratios," *Human Biology*, Vol. 14 (1941), No. 1, pp. 85–94.

10. A. Cioeco, "Variation in the sex ratios at birth in the United States," *Human Biology*, Vol. 10 (1938), pp. 35–64.
11. W. J. Martin, "A comparison of the trends of male and female mortality," *Journal of the Royal Statistical Society*, Series A, Vol. CXIV (1951), Part 3, pp. 287–298.
12. Hope Tisdale Eldridge and Jacob S. Siegel, "The changing sex ratio in the United States," *The American Journal of Sociology*, Vol. LII (1946), No. 3, pp. 224–234.
13. P. H. Karmel, "An analysis of the sources of magnitudes of inconsistencies between male and female net reproduction rates in actual populations," *Population Studies*, Vol. 2 (1948), Part 2, pp. 240–273.
14. P. H. Karmel, "The relation between male and female reproduction rates," *Population Studies*, Vol. 1 (1947), No. 3, pp. 249–274.
15. P. H. Karmel, "The relation between male and female nuptiality in a stable population," *Population Studies*, Vol. 1 (1948), No. 4, pp. 353–387.
16. J. Hajnal, "Aspects of recent trends in marriage in England and Wales," *Population Studies*, Vol. 1 (1947), No. 1, pp. 72–98.
17. J. Hajnal, "Some comments on Mr. Karmel's paper 'The relation between male and female reproduction rates'," *Population Studies*, Vol. 2 (1948), No. 3, pp. 352–360.
18. P. H. Karmel, "A rejoinder to Mr. Hajnal's comments," *Population Studies*, Vol. 2 (1948), No. 3, pp. 361–372.
19. W. S. Hocking, "The balance of the sexes in Great Britain," *Journal of the Institute of Actuaries*, Vol. 74 (1948), pp. 340–344.
20. G. Udney Yule, "A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, F.R.S.," Philosophical *Transactions of the Royal Society of London*, Series B, Vol. 213 (1925), pp. 21–87.
21. M. S. Bartlett, "Some evolutionary stochastic processes," *Journal of the Royal Statistical Society*, Series B, Vol. II (1949), pp. 211–229.
22. D. G. Kendall, "Stochastic processes and population growth," *Journal of the Royal Statistical Society*, Series B, Vol. II (1949), pp. 230–264.
23. William Feller, "Diffusion processes in genetics," *Proceedings of the Second Berkeley Symposium of Mathematical Statistics*, University of California Press (1951), pp. 227–246.
24. T. E. Harris, "Some mathematical models for branching processes," *Proceedings of the Second Berkeley Symposium of Mathematical Statistics*, University of California Press (1951), pp. 305–328.
25. P. A. P. Moran, "Estimation methods for evolutive processes," *Journal of the Royal Statistical Society*, Series B, Vol. XIII (1951), No. 1, pp. 141–146.
26. William Feller, *An Introduction to Probability Theory and its Applications*, Vol. 1, John Wiley and Sons, Inc., New York, 1950.

ESTIMATION OF VARIANCE AND COVARIANCE COMPONENTS*

C. R. Henderson

Cornell University

INTRODUCTION

The theory of variance component analysis has been discussed recently by Crump (1946, 1951) and by Eisenhart (1947). These papers and, indeed, most of the published works on estimating variance components deal with the one-way classification, with "nested" classifications, and with factorial classifications having equal subclass numbers. Also most papers on this subject are concerned with what Eisenhart (1947) has called Model II; that is, all elements of the linear model save μ are regarded as random variables. In the above cases, estimation of variance components is usually accomplished by computing the mean squares in the standard analysis of variance, equating these mean squares to their expectations, and solving for the unknown variances. These techniques are described in many statistical textbooks.

Unfortunately, research workers in some of those fields in which much use is made of variance component estimates are unable to obtain data which have the above described characteristics. This is particularly true in those fields in which survey data must be used or where, even in a well-planned experiment, the subclasses are of quite unequal size due, for example, to differences in litter numbers. Also,

*Presented at North Carolina Summer Statistics Conference June 24, 1952.

Model II is sometimes not appropriate. Instead the data more appropriately correspond to what Eisenhart called the Mixed Model. For example, the data may represent several different years, and the year effects should be regarded as fixed rather than as random variables.

It is the purpose of this paper to describe some methods for estimating variance components in the non-orthogonal case and to illustrate the methods with a small sample of butterfat records made by cows resulting from an artificial breeding program. The three methods described are:

1. Compute sums of squares as in the standard analysis of variance of corresponding orthogonal data. Equate these sums of squares to their expectations obtained under the assumption of Model II and solve for the unknown variances.
2. Obtain least squares estimates of fixed effects, "correct" the data according to these estimates of the fixed effects, and then using the corrected data in place of the original data, proceed as in Method 1.
3. Compute mean squares by a conventional least squares analysis of non-orthogonal data (method of fitting constants, weighted squares of means, e.g.). Equate these mean squares to their expectations and solve for the unknown variances.

These three methods henceforth called Method 1, Method 2, and Method 3 vary greatly in computational labor. Method 1 is the simplest. Method 2 in many cases is only slightly more difficult. Method 3 is usually much the most laborious. Method 1, however, leads to biased estimates if certain elements of the model are fixed or if some of them are correlated. Estimates obtained by Method 2 are free of the first of these biases, but not of the second. Method 3 yields unbiased estimates, but the computations required may be prohibitive. The relative sizes of the sampling variances of estimates obtained by these three methods are not known.

DESCRIPTION AND ILLUSTRATION OF METHODS OF ESTIMATION

The Data

In New York State most artificial breeding of dairy cows is accomplished with semen supplied by the New York Artificial Breeders' Cooperative, Inc. This cooperative organization has approximately 60 bulls in service. The operations of the organization are conducted in such a manner that it is largely a matter of chance to which bull's semen a particular cow is bred. This fact as well as the large number

of daughters sired by each bull make the production records of these daughters particularly suitable for studying the genetic differences among bulls and for estimating the magnitudes of other sources of variation in milk production records. Good estimates of these variances are needed in designing efficient testing and selection programs.

The difficulties in estimating the pertinent variance components are typical of those faced by research workers in animal breeding and in other fields as well. The difficulties in the present example are due to the following causes:

1. Several years' data are involved and time trends are known to be important.
2. The two major classifications of the data are sire and herd. The number of sires exceeds 100 and the number of different herds exceeds 2000.
3. The number of observations per herd-sire subclass varies; the majority being 0.

We have estimated from these data the pertinent variances. Both Method 1 and Method 2 have been employed and have yielded estimates essentially the same. A small sample of records is presented in this paper and the three methods of estimation are illustrated.

Table 1 shows the number of first lactation butterfat records in each of the year × herd × sire subclasses and also the sum of the records for each of these subclasses.

TABLE 1

Herd	Sire	Year				Total
		1	2	3	4	
1	1	3–1414	2– 981			5–2395
1	2		4–1766	2– 862		6–2628
1	3				5–1609	5–1609
2	1	1– 404	3–1270			4–1674
2	2			5–2109		5–2109
2	3			4–1563	2– 740	6–2303
3	1		3–1705			3–1705
3	2		4–2310	2–1134		6–3444
4	1	3–1113	5–1951			8–3064
4	3			3–1291	6–2457	9–3748
Total		7–2931	21–9983	16–6959	13–4806	57–24679

The Linear Model

Let y_{hijk} denote the record made in the h-th year by the k-th daughter of the j-th sire in the i-th herd. Suppose that the appropriate linear model representing these observations is

$$y_{hijk} = \mu + a_h + h_i + s_j + (hs)_{ij} + e_{hijk}$$

$h = 1, \cdots, p$ $\quad k = 1, \cdots, n_{hij}$

$i = 1, \cdots, q$ $\quad \sum_h \sum_i \sum_j n_{hij} = N$

$j = 1, \cdots, r$ $\quad$ Total number of filled subclasses $= s$

μ is common to all observations. a_h is common to all observations in the h-th year, h_i to all observations in the i-th herd, and s_j to all records made by daughters of the j-th sire; $(hs)_{ij}$ is peculiar to all records made by the daughters of the j-th sire in the i-th herd. Peculiar to each record is a random element e_{hijk} which is assumed to have mean zero and variance σ_e^2. The assumptions made concerning the other elements of the model are described for each estimation method.

Method 1*

Method 1 can be used only if it is assumed that, except for μ, all elements of the model are uncorrelated variables with means zero and variances σ_a^2, σ_h^2, σ_s^2, σ_{hs}^2, or σ_e^2. This is, of course, the Eisenhart Model II.

The following quantities are computed:

$$T = \sum_h \sum_i \sum_j \sum_k y_{hijk}^2 \qquad H = \sum_i \frac{y_{i\cdot\cdot\cdot}^2}{n_{i\cdot\cdot}}$$

$$A = \sum_h \frac{y_{h\cdot\cdot\cdot}^2}{n_{h\cdot\cdot}} \qquad S = \sum_j \frac{y_{\cdot\cdot j\cdot}^2}{n_{\cdot\cdot j}}$$

$$HS = \sum_i \sum_j \frac{y_{\cdot ij\cdot}^2}{n_{\cdot ij}} \qquad CF = \frac{y_{\cdot\cdot\cdot\cdot}^2}{N}$$

Dots in the subscripts denote summation. For example,

$$y_{h\cdot\cdot\cdot} = \sum_i \sum_j \sum_k y_{hijk} .$$

Next the expectations of the above quantities are computed. Under the assumptions of Model II, the coefficients of μ^2 and the variances in these expectations are as shown in Table 2.

*This method was first suggested to me by Dr. S. Lee Crump.

TABLE 2

	μ^2	σ_a^2	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2
T	N	N	N	N	N	N
A	N	N	K_1	K_2	K_3	p
HS	N	K_4	N	N	N	s
H	N	K_5	N	K_6	K_6	q
S	N	K_7	K_5	N	K_8	r
CF	N	K_9	K_{10}	K_{11}	K_{12}	1

N, p, q, r, s in the above table were defined in the statement of the linear model. K_1 , K_2 , $\cdots$, K_{12} must be computed as follows:

$$K_1 = \sum_h \frac{\sum_i n_{hi\cdot}^2}{n_{h\cdot\cdot}} \qquad K_7 = \sum_j \frac{\sum_h n_{h\cdot j}^2}{n_{\cdot\cdot j}}$$

$$K_2 = \sum_h \frac{\sum_j n_{h\cdot j}^2}{n_{h\cdot\cdot}} \qquad K_8 = \sum_j \frac{\sum_i n_{\cdot ij}^2}{n_{\cdot\cdot j}}$$

$$K_3 = \sum_h \frac{\sum_i \sum_j n_{hij}^2}{n_{h\cdot\cdot}} \qquad K_9 = \sum_h n_{h\cdot\cdot}^2/N$$

$$K_4 = \sum_i \sum_j \frac{\sum_h n_{hij}^2}{n_{\cdot ij}} \qquad K_{10} = \sum_i n_{\cdot i\cdot}^2/N$$

$$K_5 = \sum_i \frac{\sum_h n_{hi\cdot}^2}{n_{\cdot i\cdot}} \qquad K_{11} = \sum_j n_{\cdot\cdot j}^2/N$$

$$K_6 = \sum_i \frac{\sum_j n_{\cdot ij}^2}{n_{\cdot i\cdot}} \qquad K_{12} = \sum_i \sum_j n_{\cdot ij}^2/N$$

If the data were orthogonal, the sums of squares in the analysis of variance would be

$$\text{Among Years} = A - CF$$
$$\text{Among Herds} = H - CF$$
$$\text{Among Sires} = S - CF$$
$$\text{Herds} \times \text{Sires} = HS - H - S + CF$$
$$\text{Error} = T - A - HS + CF$$

If these same quantities are computed in spite of the non-orthogonality and are equated to their expectations, unbiased estimates of the

variances can be obtained by solving the resulting equations. The necessary expectations are derived from Table 2. To illustrate, E (Among Years) $= E(A - CF) = E(A) - E(CF)$.

Computation of the K's is facilitated by constructing from Table 1 the following two-way tables of subclass numbers (Tables 3, 4, 5).

TABLE 3

Herd	Year				Total
	1	2	3	4	
1	3	6	2	5	16
2	1	3	9	2	15
3	0	7	2	0	9
4	3	5	3	6	17
Total	7	21	16	13	57

TABLE 4

Sire	Year				Total
	1	2	3	4	
1	7	13	0	0	20
2	0	8	9	0	17
3	0	0	7	13	20
Total	7	21	16	13	57

TABLE 5

Herd	Sire			Total
	1	2	3	
1	5	6	5	16
2	4	5	6	15
3	3	6	0	9
4	8	0	9	17
Total	20	17	20	57

Also certain totals are computed from Table 1.

Year	Herd	Sire
1. 2931	1. 6632	1. 8838
2. 9983	2. 6086	2. 8181
3. 6959	3. 5149	3. 7660
4. 4806	4. 6812	
Total 24,679	Total 24,679	Total 24,679

Using the above totals and the totals in Table 1,

$$A = \frac{2931^2}{7} + \cdots + \frac{4806^2}{13} = 10{,}776{,}451$$

$$HS = \frac{2395^2}{5} + \cdots + \frac{3748^2}{9} = 10{,}970{,}369$$

$$H = \frac{6632^2}{16} + \cdots + \frac{6812^2}{17} = 10{,}893{,}666$$

$$S = \frac{8838^2}{20} + \frac{8181^2}{17} + \frac{7660^2}{20} = 10{,}776{,}278$$

$$CF = \frac{24{,}679^2}{57} = 10{,}685{,}141$$

The expectations of these quantities are presented in Table 6. The computations of these entries proceed as follows:

From Table 3, $K_1 = \dfrac{3^2 + 1^2 + 3^2}{7} + \cdots + \dfrac{5^2 + 2^2 + 6^2}{13} = 19.51$

From Table 4, $K_2 = \dfrac{7^2}{7} + \dfrac{13^2 + 8^2}{21} + \dfrac{9^2 + 7^2}{16} + \dfrac{13^2}{13} = 39.22$

From Table 1, $K_3 = \dfrac{3^2 + 1^2 + 3^2}{7} + \cdots + \dfrac{5^2 + 2^2 + 6^2}{13} = 15.10$

From Table 1, $K_4 = \dfrac{3^2 + 2^2}{5} + \cdots + \dfrac{3^2 + 6^2}{9} = 37.35$

From Table 3, $K_5 = \dfrac{3^2 + 6^2 + 2^2 + 5^2}{16} + \cdots + \dfrac{3^2 + 5^2 + 3^2 + 6^2}{17} = 21.49$

From Table 5, $K_6 = \dfrac{5^2 + 6^2 + 5^2}{16} + \cdots + \dfrac{8^2 + 9^2}{17} = 24.04$

From Table 4, $K_7 = \frac{7^2 + 13^2}{20} + \frac{8^2 + 9^2}{17} + \frac{7^2 + 13^2}{20} = 30.33$

From Table 5, $K_8 = \frac{5^2 + 4^2 + 3^2 + 8^2}{20} + \cdots + \frac{5^2 + 6^2 + 9^2}{20} = 18.51$

From Table 3, $K_9 = \frac{7^2 + 21^2 + 16^2 + 13^2}{57} = 16.05$

From Table 3, $K_{10} = \frac{16^2 + 15^2 + 9^2 + 17^2}{57} = 14.93$

From Table 4, $K_{11} = \frac{20^2 + 17^2 + 20^2}{57} = 19.11$

From Table 1, $K_{12} = \frac{5^2 + 6^2 + \cdots + 9^2}{57} = 6.19$

TABLE 6

	μ^2	σ_a^2	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2	
T	57	57.	57.	57.	57.	57	11,124,007
A	57	57.	19.51	39.22	15.10	4	10,776,451
HS	57	37.35	57.	57.	57.	10	10,970,369
H	57	21.49	57.	24.04	24.04	4	10,893,666
S	57	30.33	18.51	57.	18.51	3	10,776,278
CF	57	16.05	14.93	19.11	6.19	1	10,685,141

The equations to be solved are presented in Table 7. The first equation reads: $40.95\ \sigma_a^2 + 4.58\ \sigma_h^2 + 20.11\ \sigma_s^2 + 8.91\ \sigma_{hs}^2 + 3\ \sigma_e^2 = 91{,}310$.

TABLE 7

	σ_a^2	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2	
$A - CF$	40.95	4.58	20.11	8.91	3	91,310
$H - CF$	5.44	42.07	4.93	17.85	3	208,525
$S - CF$	14.28	3.58	37.89	12.32	2	91,137
$HS - H - S + CF$	1.58	−3.58	−4.93	20.64	4	−14,434
$T - A - HS + CF$	−21.30	−4.58	−20.11	−8.91	44	62,328

The solution to these equations is $\sigma_a^2 = 763$, $\sigma_h^2 = 4531$, $\sigma_s^2 = 1587$, $\sigma_{hs}^2 = -164$, $\sigma_e^2 = 2950$. If σ_{hs}^2 is set equal to 0, the solution is $\sigma_a^2 = 756$, $\sigma_h^2 = 4468$, $\sigma_s^2 = 1542$, $\sigma_e^2 = 2952$. These estimates, of course, have no practical value for p, q, r, and s are much too small for accurate estimation of the corresponding variances. The illustration of their computation does, however, show that even with many different classes the computations are relatively simple. We have successfully adapted most of these computations to International Business Machines operations.

A difficulty with Method 1 is that it may be inappropriate to regard the year effects as random variables. If these effects actually are fixed, the estimates of σ_h^2 , σ_s^2 , and σ_{hs}^2 are biased. The estimate of σ_e^2 may or may not be biased depending on how it is estimated. This estimate is biased if obtained from the equations of Table 7. If, however, σ_e^2 had been estimated from

$$T - \sum_h \sum_i \sum_j \frac{y_{hij\cdot}^2}{n_{hij}},$$

the within year $\times$ herd $\times$ sire subclass sum of squares, the estimate would be unbiased regardless of the assumptions concerning the a_h .

It might be well at this point to state briefly a convenient procedure for finding the expected values of quantities like H, S, etc. Substitute for the y's their corresponding linear models, and then remembering the assumptions concerning the elements of the model proceed to write out the expectations. For example,

$$E \sum_i \sum_j \frac{y_{\cdot ij\cdot}^2}{n_{\cdot ij}} = \sum_i \sum_j E \frac{y_{\cdot ij\cdot}^2}{n_{\cdot ij}}$$

$$= \sum_i \sum_j E[n_{\cdot ij}\mu + n_{1ij}a_1$$

$$+ \cdots + n_{pij}a_p + n_{\cdot ij}h_i + n_{\cdot ij}s_j + n_{\cdot ij}(hs)_{ij}$$

$$+ \sum_h \sum_k e_{hijk}]^2/n_{\cdot ij}$$

$$= \sum_i \sum_j E[n_{\cdot ij}^2\mu^2 + n_{1ij}^2a_1^2$$

$$+ \cdots + n_{pij}^2a_p^2 + n_{\cdot ij}^2h_i^2 + n_{\cdot ij}^2s_j^2 + n_{\cdot ij}^2(hs)_{ij}^2$$

$$+ \sum_h \sum_k e_{hijk}^2 + \text{cross products all having zero expectation}]/n_{\cdot ij}$$

$$= \sum_i \sum_j [n_{\cdot ij}^2 \mu^2 + n_{1ij}^2 \sigma_a^2 + \cdots + n_{pij}^2 \sigma_a^2 + n_{\cdot ij}^2 \sigma_h^2 + n_{\cdot ij}^2 \sigma_s^2 + n_{\cdot ij}^2 \sigma_{hs}^2 + \sum_h \sum_k \sigma_e^2]/n_{\cdot ij}$$

$$= N\mu^2 + \sum_i \sum_j \frac{\sum_h n_{hij}^2}{n_{\cdot ij}} \sigma_a^2 + N(\sigma_h^2 + \sigma_s^2 + \sigma_{hs}^2) + s\sigma_e^2$$

Method 2

The bias in estimating variance components due to the assumption that fixed elements of the model are random variables can be eliminated by using Method 2. At the same time the relative simplicity of Method 1 can be retained. Method 2 involves estimating the fixed effects by least squares, correcting the data in accordance with these estimates, and then applying Method 1 to the "corrected" data.

This method was used by Hazel and Terrill (1945) on data which were orthogonal except for the fixed effects. Their estimates were biased for they assumed for computational purposes that, except for fixed effects, the expectations of sums of squares of corrected data are the same as the expectations of the corresponding sums of squares of the uncorrected data. Method 2 enables one to appraise this bias and to correct for it.

Before we apply this method to our example, let us consider the general case. Suppose the linear model is

$$(1) \qquad y_a = \sum_{i=1}^{p} b_i x_{ia} + e_a \qquad a = 1, \cdots, N$$

The x's are known. The e's are uncorrelated with mean = 0 and variance = σ_e^2.

If the b's are all fixed, the least squares equations for estimating them are as shown in (2). The b's are, in fact, not all fixed in the variance components estimation problem, but they can be estimated by least squares as a matter of expediency.

$$(2) \qquad \sum_{i=1}^{p} C_{1i}\hat{b}_i = Y_1 \qquad C_{ij} = \sum_{a=1}^{N} x_{ia}x_{ja}$$

$$\sum_{i=1}^{p} C_{2i}\hat{b}_i = Y_2 \qquad Y_i = \sum_{a=1}^{N} x_{ia}y_a$$

$$\vdots \qquad\qquad \vdots$$

$$\sum_{i=1}^{p} C_{pi}\hat{b}_i = Y_p$$

It is sometimes necessary to impose one or more linear restrictions on the estimates in order to obtain a solution to equations (2).

Now suppose that $b_1, \cdots, b_s$ are fixed and also that for all $i = 1, \cdots, s$

$$E(\hat{b}_i - b_i)^2 = K_i\sigma_e^2 \tag{3}$$

It is not true that all least squares estimates have this property. For example, in our butterfat production example described in Method 1

$$E(\hat{\mu} - \mu)^2 \neq K_\mu\sigma_e^2$$

Instead

$$E(\hat{\mu} - \mu)^2 = \frac{1}{p}\sigma_a^2 + \frac{1}{q}\sigma_h^2 + \frac{1}{r}\sigma_s^2 + \lambda\sigma_{hs}^2 + K_\mu\sigma_e^2$$

Method 2 applies only to correcting data by least squares estimates for which (3) applies. It is not difficult to determine which $\hat{b}$'s qualify.

Now the data are corrected as follows (in practice only certain linear functions of the observations need to be corrected):

$$z_a = y_a - \sum_{i=1}^{s} \hat{b}_i x_{ia} \tag{4}$$

Suppose that for $i, j = s+1, \cdots, r \leq p$ all $C_{ij} = 0$ when $i \neq j$. Let

$$Z_i = \sum_a x_{ia} z_a .$$

Then compute (5). Note that

$$Z_u = Y_u - \sum_{i=1}^{s} \hat{b}_i C_{ui}$$

$$\sum_{i=s+1}^{r} Z_i^2/C_{ii} \tag{5}$$

It is found that, except for σ_e^2, the expectation of (5) is the same as the expectation of (6) with $b_1, \cdots, b_s$ assumed $= 0$.

$$\sum_{i=s+1}^{r} Y_i^2/C_{ii} \tag{6}$$

The coefficient of σ_e^2 in the expectation of (5) is increased over that of (6) by the quantity.

$$\sum_{i=1}^{s}\sum_{j=1}^{s} C^{ij}P_{ij} , \qquad \text{where} \tag{7}$$

C^{ij} are elements of the matrix inverse to the matrix of C_{ij} $(i, j = 1, \cdots, p)$, and

$$P_{uv} = \sum_{i=s+1}^{r} C_{iu}C_{iv}/C_{ii}$$

Computation of (7) is simple if s is small and if least squares equations (2) can be rewritten as (8). This can be done in many cases.

$$\sum_{j=1}^{p} C_{ij}b_j = Y_i \qquad i = 1, \cdots, s \tag{8}$$

$$\sum_{j=1}^{s} C_{ij}b_j + C_{ii}b_i = Y_i \qquad i = s+1, \cdots, p$$

Note in these equations that for all $i, j = s+1, \cdots, p$, $C_{ij} = 0$ when $i \neq j$. When equations (8) prevail, C^{ij} and $\hat{b}_i$ $(i, j = 1, \cdots, s)$ can be computed from equations (9).

$$\sum_{j=1}^{s} C'_{ij}\hat{b}_j = Y'_i \qquad (i = 1, \cdots, s) \tag{9}$$

In equations (9)

$$C'_{uv} = C_{uv} - \sum_{i=s+1}^{p} C_{iu}C_{iv}/C_{ii}$$

$$Y'_u = Y_u - \sum_{i=s+1}^{p} C_{iu}Y_i/C_{ii}$$

The least squares estimates of $b_1, \cdots, b_s$ are the solution to equations (9), and C^{ij} $(i, j = 1, \cdots, s)$ are the elements of the matrix inverse to the matrix of coefficients in (9).

Let us illustrate Method 2 with the data of Table 1. We shall now assume that the a's are fixed. First the least squares estimates of the a's are computed. This is done most simply by estimating them jointly with $d_{ij} = \mu + h_i + s_j + (hs)_{ij}$. Thus the equations are reduced to the form of equations (8). Looking at Table 1 these equations are

$$7\hat{a}_1 + 3\hat{d}_{11} + \hat{d}_{21} + 3\hat{d}_{41} = 2931$$

and similarly for the other "a" equations

$$3\hat{a}_1 + 2\hat{a}_2 + 5\hat{d}_{11} = 2395$$

and similarly for the other "d" equations
Then by (9) these equations reduce to the ones shown in Table 10.

TABLE 10

$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	
3.825	−3.825	0.	0.	− 73.500
−3.825	6.492	−2.667	0.	101.500
0.	−2.667	6.000	−3.333	41.333
0.	0.	−3.333	3.333	− 69.333

One restriction must be imposed before a solution is obtainable. A convenient one is $\hat{a}_4 = 0$. Then the solution is $\hat{a}_1 = 12.08$, $\hat{a}_2 = 31.30$, $\hat{a}_3 = 20.80$, $\hat{a}_4 = 0$.

Inverting the matrix of coefficients of Table 10 with fourth row and column deleted, the C^{ij} pertaining to the a's are obtained. These are presented in Table 11.

TABLE 11

a_1	a_2	a_3
.936438	.675000	.300000
.675000	.675000	.300000
.300000	.300000	.300000

Now the data can be corrected for the $\hat{a}$'s. For example, the corrected total for the subclass pertaining to herd 1 × sire 1 is 2395 − 3(12.08) − 2(31.30) = 2296.16. The corrected subclass and class totals are shown in Table 12.

TABLE 12

Herd	Sire			Total
	1	2	3	
1	2296.16	2461.20	1609.00	6366.36
2	1568.02	2005.00	2219.80	5792.82
3	1611.10	3277.20		4888.30
4	2871.26		3685.60	6556.86
Total	8346.54	7743.40	7514.40	23,604.34

Using the totals of Table 12 in conjunction with the subclass and class numbers of Tables 1 and 4, the corrected sums of squares are computed. Thus, $H' = (6366.36)^2/16 + \cdots + (6556.86)^2/17$. These quantities are:

$$HS' = 10{,}016{,}791 \qquad S' = 9{,}833{,}620$$

$$H' = 9{,}954{,}295 \qquad CF' = 9{,}774{,}822$$

Next the amounts by which the coefficients of σ_e^2 are increased in the corrected as compared to the uncorrected sums of squares are needed. Using (7), the P_{ij} pertaining to HS' are computed. Looking at Table 1,

$$P_{11} = \frac{3^2}{5} + \frac{1^2}{4} + \frac{3^2}{8} = 3.175$$

$$P_{12} = \frac{3(2)}{5} + \frac{1(3)}{4} + \frac{3(5)}{8} = 3.825$$

Table 13 presents the complete set of P's for HS'.

TABLE 13

a_1	a_2	a_3
3.175	3.825	0.
3.825	14.508	2.667
0.	2.667	10.000

The sum of products of corresponding entries in Table 11 and Table 13 is

$$.936438(3.175) + \cdots + .300000(10.000) = 22.53$$

Therefore, the coefficient of σ_e^2 in $E(HS') = 10 + 22.53 = 32.53$.

The P_{ij} for H' can be computed by reference to Table 3 thus

$$P_{11} = \frac{3^2}{16} + \frac{1^2}{15} + \frac{3^2}{17} = 1.159$$

$$P_{12} = \frac{3(6)}{16} + \frac{1(3)}{15} + \frac{3(5)}{17} = 2.207$$

Table 14 presents the complete set of P_{ij} for H'.

TABLE 14

a_1	a_2	a_3
1.159	2.207	1.504
2.207	9.765	4.988
1.504	4.988	6.624

Multiplying these values by those of Table 11, the addition to σ_e^2 in $E(H') = 16.54$.

The P_{ij} for S' are shown in Table 15. Referring to Table 4,

$$P_{11} = \frac{7^2}{20}, \qquad P_{12} = \frac{7(13)}{20}, \qquad \text{etc.}$$

TABLE 15

a_1	a_2	a_3
2.450	4.550	0.
4.550	12.215	4.235
0.	4.235	7.215

Then the addition to σ_e^2 in $E(S') = 2.450\ (.936438) + \cdots = 21.39$. Finally the P_{ij} for CF' are

$$P_{11} = \frac{7^2}{57}, \qquad P_{12} = \frac{7(21)}{57}, \qquad \text{etc. as shown in Table 16.}$$

TABLE 16

a_1	a_2	a_3
.860	2.579	1.965
2.579	7.737	5.895
1.965	5.895	4.491

Multiplying and summing corresponding entries of Tables 11 and 16, the addition to the coefficient of σ_e^2 in $E(CF') = 15.57$.

Table 17 shows the corrected sums of squares and their expectations,

TABLE 17

	μ^2	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2	
HS'	57.	57.	57.	57.	32.53	10,016,791
H'	57.	57.	24.04	24.04	20.54	9,954,295
S'	57.	18.51	57.	18.51	24.39	9,833,620
CF'	57.	14.93	19.11	6.19	16.57	9,774,822

The equations to be solved are presented in Table 18

TABLE 18

	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2	
$H' - CF'$	42.07	4.93	17.85	3.97	179,473
$S' - CF'$	3.58	37.89	12.32	7.82	58,798
$HS' - H' - S' + CF'$	−3.58	−4.93	20.64	4.17	3,698

The estimate of σ_e^2 can be obtained readily from the residual sum of squares after estimating the a's and d's, that is from

$$\sum_h \sum_i \sum_j \sum_k y_{hijk}^2 - \text{Reduction } (a_h , d_{ij}).$$

$$\sum_h \sum_i \sum_j \sum_k y_{hijk}^2 = 11{,}124{,}007$$

$$\text{Reduction } (a_h , d_{ij}) = 12.08\ (-73.500) + 31.30\ (101.500) + 20.80$$

$$(41.333) + HS = 3149 + 10{,}970{,}369 = 10{,}973{,}518$$

The residual sum of squares is therefore 11,124,007 − 10,973,518 = 150,489, with expectation 44 σ_e^2 . Consequently $\hat{\sigma}_e^2$ = 150,489/44 = 3,420. Substituting σ_e^2 = 3,420 in the equations of Table 18 and solving, the estimates of the variances are σ_h^2 = 3792, σ_s^2 = 409, σ_{hs}^2 = 243. It is not surprising that these estimates are different from the estimates obtained by Method 1. The sampling variances must be extremely large in both cases.

Adapting Method 2 to a model with covariates is easy to accomplish. In some problems this would simplify the computations. For example, if many years were involved, the number of fixed elements in the model could be reduced by fitting a quadratic or cubic to years instead of estimating individual yearly effects as was done in this example. If the years exhibit no trend, the simplest procedure is to regard a's as random variables and then to apply Method 1.

Method 3

When it is computationally feasible, Method 3 is the most satisfactory of the three methods for estimating variance components. For one thing it gets around the difficulty of fixed elements in the model. For another, it yields unbiased estimates even though certain elements of the model are correlated. The manner in which interference by these correlations is eliminated is described subsequently.

Unfortunately Method 3 is not likely to be computationally feasible in the non-orthogonal case unless the number of different classes is small or unless the design incorporates planned non-orthogonality and consequently the mean squares of the analysis of variance can be computed without solving least squares equations. In these two cases the expectations of the mean squares are easy to compute. For example, the analysis of the balanced incomplete block design is simple and so is the writing of the expectations of the mean squares. The basic facts needed for employing Method 3 are stated below.

Let the linear model describing y_a , the a-th observation be

$$y_a = \sum_{i=1}^{p} b_i x_{ia} + e_a \tag{10}$$

The x's are known. The e's have mean zero, are uncorrelated, and have common variance σ_e^2 . For the present we shall not specify which b's are fixed and which distributed.

Now if b_{q+1} , $\cdots$, b_p are set $= 0$, the least squares estimates of b_1 , $\cdots$, b_q are the solution to equations (11).

$$\begin{aligned} \hat{b}_1 C_{11} + \hat{b}_2 C_{12} + \cdots + \hat{b}_q C_{1q} &= Y_1 \\ \hat{b}_1 C_{21} + \hat{b}_2 C_{22} + \cdots + \hat{b}_q C_{2q} &= Y_2 \\ \text{etc.} & \end{aligned} \tag{11}$$

In equation (11)

$$C_{ij} = \sum_{a=1}^{N} x_{ia} x_{ja}$$

$$Y_i = \sum_{a=1}^{N} x_{ia} y_a$$

The reduction in sum of squares due to $\hat{b}_1$, $\cdots$, $\hat{b}_q$ is

$$R(b_1 , \cdots , b_q) = \sum_{i=1}^{q} \hat{b}_i Y_i \tag{12}$$

But since

$$\hat{b}_i = \sum_{j=1}^{q} C^{ij} Y_j ,$$

where C^{ij} are elements of the matrix inverse to the C_{ij} matrix ($i, j = 1, \cdots, q$),

$$R(b_1 , \cdots , b_q) = \sum_{i=1}^{q} \sum_{j=1}^{q} C^{ij} Y_i Y_j \tag{13}$$

Using (13), the expectation of $R(b_1 , \cdots , b_q)$ is easy to write, but not necessarily easy to compute.

$$E[R(b_1 , \cdots , b_q)] = \sum_{i=1}^{q} \sum_{j=1}^{q} C^{ij} E(Y_i Y_j) \tag{14}$$

Use will be made of the fact that (14) can be written

$$E[R(b_1 , \cdots , b_q)] = \sum_{i=1}^{q} \sum_{j=1}^{p} C_{ij} E(b_i b_j) + \sum_{i=q+1}^{p} \sum_{j=1}^{q} C_{ij} E(b_i b_j) + \sum_{i=q+1}^{p} \sum_{j=q+1}^{p} \lambda_{ij} E(b_i b_j) + q' \sigma_e^2 , \tag{15}$$

where

$$\lambda_{uv} = \sum_{i=1}^{q} \sum_{j=1}^{q} C^{ij} [C_{iu} C_{jv} + C_{iv} C_{ju}] \qquad \text{when } u \neq v,$$

and

$$\lambda_{uu} = \sum_{i=1}^{q} \sum_{j=1}^{q} C^{ij} C_{iu} C_{ju} \qquad \text{(See 14)}.$$

In most variance components problems $Eb_i b_j = 0$ $(i \neq j)$. Thus only the λ_{ii} need to be computed. q' refers to the number of independent equations in (11).

It is easy to verify that the expectation of the uncorrected total sum of squares is

$$E \sum_{a=1}^{N} y_a^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} C_{ij} E(b_i b_j) + N\sigma_e^2 . \tag{16}$$

Now it becomes clear why the residual mean square has expectation σ_e^2 regardless of the assumptions concerning the b's. Making use of (15) it is seen that

$$E[R(b_1 , \cdots , b_p)] = \sum_{i=1}^{p} \sum_{j=1}^{p} C_{ij} E(b_i b_j) + p' \sigma_e^2 . \tag{17}$$

Therefore, the expectation of the residual sum of squares is

$$(18) \qquad E[\sum_a y_a^2 - R(b_1 , \cdots , b_p)]$$

$$= \left[\sum_{i=1}^{p} \sum_{j=1}^{p} C_{ij} E(b_i b_j) + N\sigma_e^2 \right] - \left[\sum_{i=1}^{p} \sum_{j=1}^{p} C_{ij} E(b_i b_j) + p'\sigma_e^2 \right]$$

$$= (N - p')\sigma_e^2$$

Suppose that $b_{q+1} , \cdots , b_p$ are independently distributed with means = 0 and common variance σ^2. This variance can be estimated by equating $R(b_1 , \cdots , b_p) - R(b_1 , \cdots , b_q)$ to its expectation. The expectation of this difference is seen by reference to (15) and (17) to be

$$(19) \qquad \sum_{i=q+1}^{p} \sum_{j=q+1}^{p} (C_{ij} - \lambda_{ij}) E(b_i b_j) + (p' - q')\sigma_e^2$$

$$= \sum_{i=q+1}^{p} (C_{ii} - \lambda_{ii})\sigma^2 + (p' - q')\sigma_e^2$$

Then using the estimate of σ_e^2 arising from (18) an unbiased estimate of σ^2 can be obtained by equating $R(b_1 , \cdots , b_p) - R(b_1 , \cdots , b_q)$ to (19). It will be noted that the assumptions made concerning $b_1 , \cdots , b_q$ are of no consequence.

Now we shall illustrate Method 3 with our data of Table 1. If one were to carry out the usual tests of hypotheses by least squares, the following sums of squares would be computed.

Among Years = R(years, herd × sire subclasses) − R(herd × sire subclasses)

Among Herds = R(years, herds, sires) − R(years, sires)

Among Sires = R(years, herds, sires) − R(years, herds)

Herds × Sires = R(years, herd × sire subclasses) − R(years, herds, sires)

Residual = $\sum_h \sum_i \sum_j \sum_k y_{hijk}^2$ − R(years, herd × sire subclasses)

The last four of these quantities can also be used to estimate σ_h^2 , σ_s^2 , σ_{hs}^2 , and σ_e^2. If years were regarded as random variables, the first would be used to estimate σ_a^2 . Our present assumption is that the year effects are fixed, however.

According to (15) we need not be concerned with μ and the a's in the expectations since the expectation of each of the above reductions

contains the following quantity,

$$N\mu^2 + 2\sum_h n_{h\cdot\cdot}\mu a_h + \sum_h n_{h\cdot\cdot}a_h^2 .$$

This expression vanishes in the sums of squares, which are differences between two reductions.

Aside from μ and a_h terms, the expectations of the pertinent reductions are those shown in Table 19.

TABLE 19

	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2
$\sum_h\sum_i\sum_j\sum_k \gamma_{hijk}^2$	N	N	N	N
R(years, herd × sire subclasses)	N	N	N	$p+s-1$
R(years, herds, sires)	N	N	K_1	$p+q+r-2$
R(years, herds)	N	K_2	K_3	$p+q-1$
R(years, sires)	K_4	N	K_5	$p+r-1$

The entries other than the K's in Table 19 are derived from (15) and (16). The K's are computed by (14). The expectations of the sums of squares of the analysis of variance are presented in Table 20.

TABLE 20

S.S.	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2
Among Herds	$N-K_4$	0	K_1-K_5	$q-1$
Among Sires	0	$N-K_2$	K_1-K_3	$r-1$
Herds × Sires	0	0	$N-K_1$	$s-q-r+1$
Residual	0	0	0	$N-p-s+1$

The computations of the various reductions proceed as follows in our example. R(years, herd × sire subclasses) = 10,973,518 as was shown in the description of Method 2. To obtain R(years, herds, sires) the equations of Table 21 need to be solved. In these equations μ is estimated jointly with each a_h, while h_4 and s_3 are set = 0. These three or some other set of three restrictions on the estimates are necessary.

TABLE 21

	a_1	a_2	a_3	a_4	h_1	h_2	h_3	s_1	s_2	
a_1	7	0	0	0	3	1	0	7	0	2931
a_2	0	21	0	0	6	3	7	13	8	9983
a_3	0	0	16	0	2	9	2	0	9	6959
a_4	0	0	0	13	5	2	0	0	0	4806
h_1	3	6	2	5	16	0	0	5	6	6632
h_2	1	3	9	2	0	15	0	4	5	6086
h_3	0	7	2	0	0	0	9	3	6	5149
s_1	7	13	0	0	5	4	3	20	0	8838
s_2	0	8	9	0	6	5	6	0	17	8181

The solution is

$$a_1 = 414.77 \qquad h_1 = 6.13 \qquad s_1 = 2.48$$
$$a_2 = 419.83 \qquad h_2 = -8.15 \qquad s_2 = 15.09$$
$$a_3 = 412.39 \qquad h_3 = 143.05$$
$$a_4 = 368.59$$

$$R(\text{years, herds, sires}) = 414.77(2931) + \cdots + 15.09(8181)$$
$$= 10{,}921{,}107$$

In order to compute $R(\text{years, herds})$ the s_1 and s_2 rows and columns of Table 21 are deleted and the resulting equations solved. The solution is

$$a_1 = 414.76 \qquad h_1 = 11.35$$
$$a_2 = 422.99 \qquad h_2 = -6.35$$
$$a_3 = 418.32 \qquad h_3 = 150.16$$
$$a_4 = 366.30$$

$$R(\text{years, herds}) = 414.76(2931) + \cdots + 150.16(5149)$$
$$= 10{,}919{,}698$$

The reduction due to years and sires requires solution of the equations of Table 21 with the h_1 , h_2 , h_3 rows and columns deleted. The solution is

$$a_1 = 425.46 \qquad a_3 = 407.72$$
$$a_2 = 461.13 \qquad a_4 = 369.69$$

$$s_1 = -6.75 \qquad s_2 = 48.38$$

$$R(\text{years, sires}) = 425.46(2931) + \cdots + 48.38(8181)$$

$$= 10{,}800{,}679$$

The computations of the K's in Table 19 require inversions of certain matrices. To obtain K_1, the inverse of the matrix of coefficients in Table 21 is needed. This inverse matrix is presented in Table 22. The entries to the left of the diagonal are omitted since the matrix is symmetric.

TABLE 22

	a_1	a_2	a_3	a_4	h_1	h_2	h_3	s_1	s_2
a_1	.64297	.41685	.15893	.03469	−.07895	−.02813	−.05580	−.46226	−.22447
a_2		.42381	.16668	.03183	−.06608	−.04169	−.09296	−.38257	−.21929
a_3			.19176	.02719	−.03647	−.08558	−.04440	−.13108	−.12625
a_4				.10925	−.06501	−.04759	−.04686	−.00004	.02410
h_1					.14349	.06382	.09075	.00834	−.05104
h_2						.14976	.07769	−.02062	−.02907
h_3							.23943	.00581	−.07214
s_1								.46163	.25050
s_2									.28088

Now we need the coefficients of σ^2_{hs} in the expectations of squares and products of the right members of the equations of Table 21. These computations are facilitated by setting up Table 23.

TABLE 23

Herd × sire subclasses

Right members	11	12	13	21	22	23	31	32	41	43
$y_{1..}$	3	0	0	1	0	0	0	0	3	0
$y_{2...}$	2	4	0	3	0	0	3	4	5	0
$y_{3...}$	0	2	0	0	5	4	0	2	0	3
$y_{4...}$	0	0	5	0	0	2	0	0	0	6
$y_{.1..}$	5	6	5	0	0	0	0	0	0	0
$y_{.2..}$	0	0	0	4	5	6	0	0	0	0
$y_{.3..}$	0	0	0	0	0	0	3	6	0	0
$y_{..1.}$	5	0	0	4	0	0	3	0	8	0
$y_{..2.}$	0	6	0	0	5	0	0	6	0	0

The coefficients of σ^2_{hs} in the squares and products of right members are the squares or products of appropriate rows in Table 23. For example, the coefficient of σ^2_{1s} in $E\ y_1^2...$ is $3^2 + 1^2 + 3^2 = 19$. That in $E(y_1...y_2...)$ is $3(2) + 1(3) + 3(5) = 24$. The complete set of coefficients is presented in Table 24, with entries to the left of the diagonal omitted due to the symmetry of the matrix.

TABLE 24

	a_1	a_2	a_3	a_4	h_1	h_2	h_3	s_1	s_2
a_1	19	24	0	0	15	4	0	43	0
a_2		79	16	0	34	12	33	71	48
a_3			58	26	12	49	12	0	49
a_4				65	25	12	0	0	0
h_1					86	0	0	25	36
h_2						77	0	16	25
h_3							45	9	36
s_1								114	0
s_2									97

Multiplying and summing corresponding entries of Tables 22 and 24,

$$K_1 = 19(.64297) + 2(24)(.41685) + \cdots + 97(.28088)$$

$$= 38.29$$

Calculation of K_2 requires the inverse of the matrix of coefficients of Table 21 with s_1 and s_2 columns and rows deleted. This inverse is presented in Table 25.

TABLE 25

	a_1	a_2	a_3	a_4	h_1	h_2	h_3
a_1	.17524	.03588	.03770	.03027	−.06049	−.04552	−.03629
a_2		.10581	.05361	.03375	−.06365	−.06022	−.09421
a_3			.13358	.03635	−.05523	−.09823	−.07138
a_4				.10523	−.05576	−.04461	−.03433
h_1					.12204	.05734	.06178
h_2						.14663	.06867
h_3							.20025

Also needed in the computation of K_2 are the coefficients of σ_s^2 in the expectations of squares and products of right members of the least squares equations. Table 26 facilitates this computation.

TABLE 26

Right members	Sires		
	1	2	3
$y_{1...}$	7	0	0
$y_{2...}$	13	8	0
$y_{3...}$	0	9	7
$y_{4...}$	0	0	13
$y_{.1..}$	5	6	5
$y_{.2..}$	4	5	6
$y_{.3..}$	3	6	0

The coefficient of σ_s^2 in $E(y_1^2...)$ is $7^2 = 49$, in $E(y_1...y_2...)$ is $7(13) = 91$, etc. The complete set is shown in Table 27.

TABLE 27

	a_1	a_2	a_3	a_4	h_1	h_2	h_3
a_1	49	91	0	0	35	28	21
a_2		233	72	0	113	92	87
a_3			130	91	89	87	54
a_4				169	65	78	0
h_1					86	80	51
h_2						77	42
h_3							45

Now $K_2 = 49\ (.17524) + 2(91)\ (.03588) + \cdots + 45\ (.20025)$

$= 42.29$

K_3 is obtained from Table 24 and 25, thus

$K_3 = 19\ (.17524) + 2\ (24)\ (.03588) + \cdots + 45\ (.20025)$

$= 31.71.$

In a similar manner K_4 is found to be 22.57 and K_5 to be 22.57 (the equality of K_4 and K_5 is only a coincidence.)

Table 28 presents the pertinent reductions and their expectations (excluding μ and a_h terms)

TABLE 28

	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2	
$\Sigma\Sigma\Sigma\Sigma\ y_{hijk}^2$	57	57	57	57	11,124,007
R(years, herd × sire subclasses)	57	57	57	13	10,973,518
R(years, herds, sires)	57	57	38.29	9	10,921,107
R(years, herds)	57	42.29	31.71	7	10,919,698
R(years, sires)	22.57	57	22.57	6	10,800,679

Then the equations to be solved are those of Table 29.

TABLE 29

	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2	
Among Herds	34.43	0	15.72	3	120,428
Among Sires	0	14.71	6.58	2	1,409
Herds × Sires	0	0	18.71	4	52,411
Residual	0	0	0	44	150,489

The solution to these equations is $\sigma_h^2 = 2255$, $\sigma_s^2 = -1295$, $\sigma_{hs}^2 = 2070$, and $\sigma_e^2 = 3420$.

ESTIMATION OF COMPONENTS OF COVARIANCE

The same general principles described in Methods 1, 2, and 3 for estimating variances can be employed to estimate covariances. To illustrate, suppose an observation is made on each of the progeny resulting from single crosses among inbred lines. If y_{ijk} is the observation on the k-th progeny of the i-th male line by the j-th female line cross, a model which might reasonably be assumed is

$$y_{ijk} = \mu + g_i + g_j + m_j + s_{ij} + e_{ijk}$$

where g_i is the general combining ability of the i-th line, g_j is the general combining ability of the j-th line, m_j is the maternal ability (exclusive of the genes transmitted to the progeny) of the j-th line, s_{ij} is peculiar to crosses $i \times j$ and of $j \times i$, and e_{ijk} is a random error. Suppose further that the elements of the model fit Eisenhart's Model II except that $E(g_i m_i) = \sigma_{gm}$.

The problem is to estimate the variances and σ_{gm}. If Method 3 is used, unbiased estimates of the variances are obtained. If Method 1 is used, the estimates are biased due to the presence of σ_{gm}. If the least squares estimates of g_i and m_j are computed, an unbiased estimate of σ_{gm} can be derived from $\sum_i \hat{g}_i \hat{m}_i$, the expectation of which is $(p-1)\sigma_{gm} + \sum_i k_i \sigma_e^2$, where p is the number of lines and $k_i \sigma_e^2$ is the covariance between $\hat{g}_i$ and $\hat{m}_i$, assuming that g_i and m_i are fixed.

A more frequently occurring type of covariance estimation problem in animal breeding arises in connection with estimation of genetic and phenotypic correlations. Observations are taken on two or more traits in some population. The following linear models characterize such observations on two traits.

(20)
$$y_a = \sum_{i=1}^{p} b_i x_{ia} + e_a$$

$$y'_a = \sum_{i=1}^{p} b'_i x_{ia} + e'_a$$

b_i and b'_i are fixed for $i = 1, \cdots, q$

b_i and b'_i are random variables for $i = q+1, \cdots, p$

So far as the random variables are concerned,

$$Eb_i = Eb'_i = 0 \qquad E(e'_a)^2 = \sigma_{e'}^2$$

$$E(b_i)^2 = \sigma_i^2 \qquad E(b_i b'_i) = \sigma_{ii'}$$

$$E(b'_i)^2 = \sigma_{i'}^2 \qquad E(e_a e'_a) = \sigma_{ee'}$$

$$E(e_a)^2 = \sigma_e^2 \qquad \text{All other covariances} = 0$$

Now in place of least squares reductions like

$$\sum_i \hat{b}_i Y_i \qquad \text{or} \qquad \sum_i Y_i^2/C_{ii}$$

we substitute

$$\sum_i \hat{b}_{i'} Y_i \qquad \text{or} \qquad \sum_i Y_i Y_{i'}/C_{ii},$$

where $Y_{i'}$, refers to a right member of the least squares equations for the second trait.

Then the expectations of these reductions are exactly the same as described for estimation of variance components except that $\sigma_{ii'}$, is substituted for σ_i^2 and $\sigma_{ee'}$, is substituted for σ_e^2. Therefore, any of the

three methods for estimating variances can be used equally well for estimating covariances when (20) is the model.

REFERENCES

Crump, S. L. The estimation of variance components in analysis of variance. *Biom. 2*: 7–11, 1946.

Crump, S. L. The present status of variance component analysis. *Biom. 7*: 1–16, 1951.

Eisenhart, C. The assumptions underlying analysis of variance. *Biom. 3*: 1–21, 1947.

Hazel, L. N. and C. E. Terrill. *Heritability of weaning weight and staple length in range Rambouillet lambs*, 1945.

QUERIES

GEORGE W. SNEDECOR, *Editor*

100 QUERY: We have been using the technique given in Snedecor's "Statistical Methods", section 11:10 and have come across some results which have raised a question of correct procedure. The data are the weights (grams) of foetal membranes in swine of the 25th. day of gestation.

Mating	Litter Size			
	≤9		>9	
	k_1	$\bar{x}_1$	k_2	$\bar{x}_2$
CC	2	12.700	4	9.525
PC	4	8.575	1	5.100
CP	2	10.600	3	7.200
PP	3	8.233	3	8.400

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Litter Size	1	11.06	
Mating	3	24.82	
Error	14	108.23	7.73

Our calculations yield the following results:

$$\sum WD^2 = 37.03$$

$$(\sum WD)^2/\sum W = 24.28$$

$$\text{Interaction} \qquad 12.75$$

Since interaction may be assumed negligible, we proceeded as follows:

Litter Size Sum of Squares, unadjusted	=	11.06
Litter Size Sum of Squares, adjusted	=	24.28
Correction for Disproportion	=	−13.22

Is it possible to have a negative correction? If so, is it correct to use algebraic subtraction in this next step?

Mating Sum of Squares, unadjusted	=	24.82
Correction for Disproportion	=	−13.22
Difference for Adjusted Sum of Squares	=	38.04

ANSWER: For the 2 × 2 table it is possible to have a negative correction to be subtracted algebraically as you have done. That is, the adjusted sum of squares may be larger than the unadjusted.

For the $R \times 2$ table, the method of adjustment given in table 11.22 of my text is incorrect. My attention was called to this recently by Dr. J. O. Irwin. This adjusting device applies to the 2 × 2 table but not to the larger one.

A method for calculating the adjusted sum of squares for the rows is illustrated below. It is similar to the scheme for calculating the interaction in table 11.22. The distinction is that the sums, $S = \bar{x}_1 + \bar{x}_2$, are used instead of the differences.

$W = \frac{k_1 k_2}{k_1 + k_2}$	$S = \bar{x}_1 + \bar{x}_2$	WS
1.3333	22.225	29.6326
0.8000	13.675	10.9400
1.2000	17.800	21.3600
1.5000	16.633	24.9495
$\Sigma W = 4.8333$		$\Sigma WS = 86.8821$

$\sum WS^2$	=	1,603.38
$(\sum WS)^2 / \sum W$	=	1,561.77
Sum of Squares for mating	=	41.61

The final analysis of variance for your data is in this form (I carried more decimals than you did).

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Litter Size	1	24.32	24.32
Mating	3	41.61	13.87
Interaction	3	12.69	4.23
Error	14	108.23	7.73

101 QUERY: In three experiments I have observed that treatment A produces more than B, but in no case was the difference significant. The experiments differed in design so that I cannot combine the original observations. Yet I think there should be some way to take advantage of the fact that in every experiment the probability was less than 0.2. It doesn't seem that this would be likely to happen if there were no real difference in yield.

ANSWER: R. A. Fisher devised a method for combining comparable probabilities such as you seem to have. See his "Statistical Methods for Research Workers", section 21.1. Bancroft and Anderson in their "Statistical Theory in Research", section 12.6, have emphasized the allowable variation in design.

The method depends on the facts that (i) the sum of several values of chi-square is itself distributed as chi-square with the appropriate number of degrees of freedom and that (ii) minus twice the natural logarithm of a probability is distributed as chi-square with 2 degrees of freedom.

As an example I shall parallel Fisher's illustration, using common logarithms and changing the probabilities to conform to your specification.

P	$-\log P$	Degrees of Freedom
0.145	0.8386	2
0.200	0.6990	2
0.087	1.0605	2
Total	2.5981	6

$\chi^2 = 2(2.3026)(2.5981) = 11.965$

The factor 2.3026 changes the common logarithm to the natural.

For 6 degrees of freedom, $\chi^2 = 11.965$ corresponds to the probability of about 0.062. Thus, the evidence against the null hypothesis is greater than that in any of your individual experiments.

102 QUERY: I set up a $2^3 \times 4$ factorial experiment to study the effect of light on the feathering of chicks. I planned to use the second and third order interactions to estimate error, but the four birds in one lot died from causes not associated with the treatment. Is there any way to analyze the variance of the remaining treatments? The data enclosed are percentages of undesirable feathers at 10 and 12 weeks of age.

ANSWER: R. L. Anderson suggested a way to analyze the variance of an unreplicated experiment with a missing treatment. "If high-order interactions are used as estimates of error, missing values should be determined by the process of minimizing the error variance." Biometrics Bulletin, Volume 2, page 43, 1946.

For illustration, it will be sufficient to use only part of your data. I have chosen the two levels of starting light treatment, s, (10 and 15 hours per day during the first six weeks), the two dates of killing, k, (10 and 12 weeks), and three of the levels of finishing light treatment, f, (12, 18 and 24 hours from six weeks to killing). Again, for illustration only, I shall assign to the estimate of error not only the 2 degrees of freedom for the three-factor interaction, SKF, but also the 2 for the interactions of deviations from linear trend with the 2-level treatments, SF_Q and KF_Q .

The following work sheet contains the data for females, among which the missing datum occurred.

The sum of the four indicated squares is $381.85 - 23.62x + 0.5x^2$. This is minimized by $x = 23.62$. For those who are not familiar with the rule for differentiation, use the formula,

$$x = \frac{-(\text{coefficient of } x)}{2(\text{coefficient of } x^2)}$$

The substitution of $x = 23.62$ in the sum of squares for error yields

Sum of Squares for Error	= 102.90
Mean Square (4 − 1 = 3 d.f.)	= 34.30

The degrees of freedom for error are, as usual, decreased by one to compensate for the missing datum.

Effects	s_1						s_2						Differences	Divisor	$\frac{(\text{Difference})^2}{\text{Divisor}}$
	k_1			k_2			k_1			k_2					
	f_1	f_2	f_3	f_1	f_2	f_3	f_1	f_2	f_3	f_1	f_2	f_3			
	17.6	32.1	9.0	16.6	20.1	20.9	18.8	x	9.5	13.9	24.0	17.6			
S	−	−	−	−	−	−	+	+	+	+	+	+			
K	−	−	−	+	+	+	−	−	−	+	+	+			
SK	+	+	+	−	−	−	−	−	−	+	+	+			
F_L	−		+	−		+	−		+	−		+			
F_Q	−	+2	−	−	+2	−	−	+2	−	−	+2	−			
SF_L	+		−	+		−	−		+	−		+			
KF_L	+		−	−		+	+		−	−		+			
SF_Q	+	−2	+	+	−2	+	−	+2	−	−	+2	−	$-52.1 + 2x$	24	$\frac{2714.41 - 208.4x + 4x^2}{24}$
KF_Q	+	−2	+	−	+2	−	+	−2	+	−	+2	−	$9.9 - 2x$	24	$\frac{98.01 - 39.6x + 4x^2}{24}$
SKF_L	−		+	+		−	+		−	−		+	0.1	8	$\frac{0.01}{8}$
SKF_Q	−	+2	−	+	−2	+	+	−2	+	−	+2	−	$79.7 - 2x$	24	$\frac{6352.09 - 318.8x + 4x^2}{24}$

The value $x = 23.62$ is now substituted in the difference for each effect not assigned to error and the mean square calculated in the regular way.

Professor Anderson writes as follows: "As you indicate, the treatment comparisons will be slightly biased if the usual analysis of variance is used with the missing value. One might suggest using the covariance technique I used in my Biometrics Bulletin article on 'Missing Plot Techniques.' But since each comparison has a single degree of freedom, it might be easier to compute the variance for each comparison (hence adjust the variance instead of the numerator). However even this would be so complicated that I doubt if anyone would bother to adjust for the bias."

ABSTRACTS

ENAR Joint Meeting of The Biometric Society and The Institute of Mathematical Statistics, April 29–May 1, 1953.

212 J. EDWARD JACKSON AND ROBERT H. MORRIS. (Eastman Kodak Company, Rochester, New York). **The Application of Multivariate Quality Control to a Photographic Problem.**

The use of univariate Quality Control procedures in photographic problems fails in many cases because several highly correlated measurements must be studied simultaneously. The use of Hotelling's T^2 in the form $T^2 = (x - x_0)S^{-1}(x - x_0)'$ leads to practical difficulties because the determinant of S is usually quite small. However, the use of Principal Components in conjunction with the T^2 technique offers a workable control tool which, in addition, supplies better photographic measurements than those previously used. A practical example will be given illustrating the use of this technique on an actual process control problem.

213 H. C. BATSON. (University of Illinois College of Medicine, Chicago). **Factorial Chi-square Analysis of Data from Experiments in Immunology.**

The previously unpublished method developed by A. E. Brandt for Chi-square analysis of attribute data from factorial experiments is presented in sufficient detail to permit employment of the technique by others. The method, designated "Factorial Chi-square" by Brandt, is essentially a form of analysis of variance employing normalized estimates of variance. The error term used in variance ratios is the normalized population variance calculated from the totals of the outcome groups in each experiment. Factorial coefficients are used directly in calculating Chi-square values based on individual degrees of freedom. Applications of the method are illustrated by means of examples taken from experimental immunology. Included among the examples are a 2×2 factorial, a $2 \times 2 \times 3$ factorial, a 2×2 factorial with 4-fold replication and a single classification experiment in which the number of individuals per experimental unit are unequal.

214 IRWIN D. J. BROSS. (Cornell University Medical College). **Applications of Non-Parametric Methods to Medical Data.**

Four rank order significance tests are briefly described with the aid of a Drion diagram (Ann. Math. Stat., 1952, pp. 653–74). Pros and cons of the methods as applied to medical data are considered. Three illustrations are given of situations where the methods seem especially appropriate. In particular it is shown that Drion's small sample version of the Smirnov test can be used to construct self-stopping follow-up schemes that would appear to be of value in clinical trials.

215 DAVID G. KENDALL. (Magdalen College, Oxford and Princeton University). **Stochastic Growth and Mutation Processes.**

The paper presents a simple technique (worked out in collaboration with W. A. O'N. Waugh) for dealing with the Bellman-Harris integral equation for stochastic growth, when the division-time distribution, $dG(\tau)$, is a weighted convolution of χ_2^2-distributions. The method is then applied to the formulation and study of a stochastic model of bacterial mutation incorporating "phenotypic delay".

216 GEORGE E. P. BOX AND W. A. HAY. (Institute of Statistics, Raleigh, N. C. and Imperial Chemical Industries, Manchester, England). **Statistical Designs for the Efficient Removal of Trends Occurring in Comparative Experiments with Applications in Biological Assay.**

Consider the regression of a quantitative result y, which is measurable but subject to error, on a quantitative factor x, which can be fixed by the experimenter at any desired level. For example in biological assay y would be biological response and x the dose of drug. A design is derived by means of which two such regressions (corresponding for example to a test preparation and a standard preparation of a drug) may be compared whilst eliminating, without loss of efficiency, a smooth time trend in the response y, which occurs due to factors outside the control of the experimenter. The experiments are performed in k pairs. In the u-th pair a dose d_u of the test preparation and a dose αd_u of the standard preparation is administered, where $u = 1, 2, \cdots, k$ and α is a suitably chosen constant so that roughly the same response is obtained with test and standard preparations. The doses d_u are such that $d_u, d_u^2, \cdots, d_u^q$ are orthogonal with the elements $t, t^2, \cdots, t^p$ of a polynomial in time fitted to the pair means, and q and p are so chosen that the polynomials adequately represent the dose and time effects. In the important special case where a linear time trend is adequate the device of angular random-

isation (Box, G. E. P. "Multi-factor Designs of First Order", Biometrika, Vol. 39 (1952), pp. 49–57) may be used to select the design. This ensures that subsequent normal theory statistical tests applied are exact whatever the distribution of the response and whether the mathematical model can be strictly justified or not.

A method is indicated whereby residual time effects within pairs may be eliminated using the between pairs trend estimate.

217 W. S. CONNOR AND W. H. CLATWORTHY. (National Bureau of Standards). **Necessary Conditions for the Existence of Partially Balanced Incomplete Block Designs with Two Associate Classes.**

For a partially balanced incomplete block design with two associate classes and with parameters $v, b, r, k, n_1, n_2, \lambda_1, \lambda_2$, and p_{jk}^i $(i,j,k = 1,2)$, the following theorem has been proved: If (i) $v > b$, then it is necessary that (a) Δ be a perfect square and (b) either $r - r_1 = 0$, or $r - r_2 = 0$; (ii) $v = b$ and v is even, then it is necessary that (a) Δ be a perfect square and (b) $r - r_u$ be a perfect square when a_u is odd $(u = 1, 2)$; (iii) $v = b$, v is of the form $4t + 3$ $(t = 0, 1, 2, \cdots)$, and Δ is not a perfect square, then it is necessary that $(r - r_1)(r - r_2)$ be a perfect square, and (iv) $v < b$ and v is even, then it is necessary that Δ be a perfect square where $r_u = \frac{1}{2}[(\lambda_1 - \lambda_2)(-\gamma + (-)^u\sqrt{\Delta}) + (\lambda_1 + \lambda_2)]$, $(u = 1, 2)$, $\gamma = p_{12}^2 - p_{12}^1$, $\Delta = \gamma^2 + 2\beta + 1$, $\beta = p_{12}^1 + p_{12}^2$, and a_1 and a_2 are nonnegative integers such that $a_1 + a_2 = v - 1$. Examples are given of sets of parameters which fail to satisfy these conditions.

218 A. C. COHEN, JR. (The University of Georgia, Athens). **Estimation in Truncated Bivariate Normal Distributions (Preliminary Report).**

Maximum likelihood estimators of the parameters of a bivariate normal population are developed for samples which are subjected to a truncation on one of the variates at known terminals. Both single and double truncations with the number of missing (unmeasured) observations either known or unknown are considered. Asymptotic variances of the estimates are obtained from the likelihood information matrices.

219 R. F. DRENICK AND P. NESBEDA. (RCA Victor Division, Camden, N. J.). **On a Class of Optimum Linear Predictors.**

Prediction is the problem of projecting into the future a set of observed data in order to obtain estimate for future observable data. For optimum prediction one assigns, through some considerations which are

not part of the method, a loss function representing the penalty for error. An optimum prediction procedure is the one which minimizes, in the long run, this penalty. N. Wiener pointed out that the optimum mean square predictor is linear if the interference affecting the observations has gaussian probability distribution. By using a method of estimation due to Pitman (*Estimation of location and scale parameters.* Biometrika 30 (1939)) the authors show that the class of linear predictors is characterized by the gaussian probability distribution and by a loss function more general than rms, namely, one which is symmetric and has continuous derivatives. Most of the loss functions of practical interest are in this category. Furthermore any such loss function leads to the same linear predictor X_p which has also the property: $P(X_p - x \leq k) = \max$ for all $k > 0$, x being the true value. (Work sponsored by the Bureau of Aeronautics.)

D. B. DUNCAN. (Virginia Polytechnic Institute, Blacksburg).

220 Multiple Range Tests and the Multiple Comparisons Test (Preliminary Report).

Several methods are available for testing differences between treatments in an analysis of variance. The two considered most satisfactory are one by Newman (1939) and Keuls (1952) and the Multiple Comparisons Test by Duncan (1951). Both employ repeated homogeneity tests. The Newman-Keuls test is simpler because it uses repeated range tests instead of F tests as used by the Multiple Comparisons Test. The latter is generally more sensitive owing partly to this reason but mostly to the relaxation of the significance levels of some of the tests considered to be of diminished importance. This paper presents: a new Multiple Range Test which achieves the simplicity of the Newman-Keuls test by using range tests and most of the sensitivity of the Multiple Comparisons Test by using the special significance levels, and an improved set of application rules for the Multiple Comparisons Test. Each of these is recommended for use depending on the relative needs for simplicity or sensitivity. The special system of significance levels is discussed in some detail. The author is indebted to W. Beyer in the determination of significant ranges for the new test which is still in progress.

NORMAN LLOYD JOHNSON. (Institute of Statistics, Chapel Hill, N. C.).

221 Sequential Procedures in Component of Variance Problems.

Three types of sequential procedure are discussed. (1) Procedures based on the use of sample variance at each stage, (2) procedures based

on successive independent groups of samples, and (3) procedures based on the isolation of independent degrees of freedom at each stage.

In cases (2) and (3) approximate formulae for the average sampling number (ASN) are evaluated. It is found that (3) gives a lower ASN than (2) when the successive groups in (2) are of small size but that (2) gives the lower ASN (formally) for larger sized groups. It is conjectured that the latter (ASN for large group size in (2)) may be related to the ASN under procedure (1).

222 MARVIN ZELEN. (National Bureau of Standards). **The Analysis of Some Incomplete Block Designs with a Missing Block.**

During the course of carrying out an experiment the situation may arise where all the observations from a particular block are missing. This paper outlines the statistical analysis if a block is missing for (a) balanced incomplete block designs, (b) partially balanced incomplete block designs of the group divisible type, (c) simply linked block designs, and (d) several miscellaneous cases which depend on the relationship of the treatments in the missing block to each other.

223 C. I. BLISS, THEODORE GREINER AND HARRY GOLD. (Yale University and Cornell University Medical College). **Estimating the Dose of a Cardiac Glycoside for Human Subjects.**

The threshold dose of digitoxin has been estimated for each of 89 human subjects, ranging in age from 3 to 74 years. When dosage was expressed in micrograms per pound of body weight, children required a 59% larger dose than adults. Four partial regression equations have been computed from the same data, for children and adults separating males and females, in each equation relating the logarithm of the threshold dose per individual to his log-weight, log-height and age. In every case, the regression on age was less than its error. Moreover, the regression equations of log-dose upon log-weight and log-height did not differ significantly between children and adults in either sex. Combining age groups within each sex, log-height contributed no information and the regressions on log-weight for males and for females could be fitted by parallel lines with a slope of .55, with males requiring 27% more digitoxin than females. Adjusting dosage approximately to the square root of the body weight should correct adequately for variations in the age and size of patients.

CLYDE Y. KRAMER AND DAVID B. DUNCAN. (Virginia Polytechnic Institute, Blacksburg). **224 On the Analysis of Variance of a Two-Way Classification with Unequal Subclass Numbers.**

This paper reviews important current methods on the analysis of variance of a two-way classification with unequal subclass numbers, presenting them from a unified point of view and also proposes a new method which has special advantages for particular situations.

For cases in which no interaction is present, the most efficient procedure is the method of fitting constants. The method of weighted squares of means is much simpler to apply but would generally sacrifice too much efficiency. The virtue of the new method is that it is equally simple for such cases and is more efficient than the method of weighted squares of means.

There are situations in which this would not be true and in which case the method of weighted squares of means should be used. A quick test for recognizing situations of this type is provided.

The new method tends to give weight to subclass means more in proportion to the numbers on which they are based than does the method of weighted squares of means. When the number of subclasses are large the time and effort saved by the proposed method may outweigh the loss of efficiency.

ALBERT B. PARKS. (Bureau of Human Nutrition and Home Economics, U.S.D.A.). **225 Ranking Versus Scoring in Palatability Tests Using Small, Trained Panels.**

A design for the simultaneous scoring and ranking of samples in palatability tests is presented and illustrated by an experiment with a small, trained panel used in the detection of off-flavors. Parallel sets of results, in the form of mean scores as compared with estimated treatment ratings based on ranks, are discussed in terms of validity, consistency of tests of significance, relative discrimination among treatments, and performance of panel members. It is suggested that this design may be useful in testing a ranking technique when a series of experiments is in progress and a scoring method has been employed. Continuity of past results is maintained.

MAX A. BERSHAD, WILLIAM N. HURWITZ AND RALPH S. WOODRUFF. (Bureau of the Census). **226 Sampling for Time Series.**

This presentation examines a rotation pattern for a sample which is a composite of the optimum rotation for month-to-month change and the

optimum for totals (or averages) over a time period. In connection with this rotation pattern, an estimate, which is a composite of a chained estimate and a simple unbiased estimate, is examined. The variances of some common statistics are presented for the composite estimate and compared with the variances for the chained and unbiased estimates. The composite estimate is proved to be the best linear estimate of level for any one time period when all past observations, first assumed as numerous, are taken into account. The form of the estimate when many observations are not available is given and is indicated as approximated well by the formula for many observations.

227 WALTER A. HENDRICKS. (Bureau of Agricultural Economics). **Response Rates and Selectivity in Mail Surveys.**

Cumulative means per sampling unit from successive mailed returns often follow a simple equation of the form, $Y = aX^b$. This provides a simple basis for extrapolating estimates to a 100 percent-return equivalent. Response rates seem to follow a law identical to relationships found in dosage-mortality studies in biology. These relationships seem to hold when surveys are conducted on a particular subject in a restricted universe. When general-purpose surveys are conducted, the relationship between percentage returns and means per sampling unit for individual items becomes more complex. In general, there seem to be factors which induce a respondent to fill out a mailed questionnaire and other factors which work simultaneously in the opposite direction. The net effect of these opposing influences governs the response rate to a mailed survey. The particular way in which the items to be estimated from the survey are correlated with the factors affecting response determines the form of the relationship between cumulative means and cumulative response rates from repeated mailings. In general-purpose surveys there does not seem to be any easily predictable pattern because of the multiplicity of interrelationships.

French Region, February 25, 1953.

228 J. SUTTER AND L. TABAH. **Resultats au Test Mosaique de Gille dans 1.244 Fratries de Deux Enfants et 380 Couples de Jumeaux.**

Dans les fratries de deux enfants, après réduction du facteur âge, les performances des garçons précédés ou suivis d'un garçon, apparaissent significativement inférieures à celles des garçons précédés ou suivis d'une fille. Par contre, les filles ne sont pas influencées par la composition

de la fratrie. Il existe une corrélation positive très significative entre les performances au test dans les fratries et la durée de l'intervalle intergénésique. Le rang de naissance n'exerce pas d'influence dans les diverses catégories de fratries, sauf dans les fratries de deux garçons ou deux filles, urbaines et dont l'intervalle intergénésique est faible; les premiers nés apparaissent alors inférieurs aux deuxièmes nés. Le coefficient de corrélation intra-classe apparaît de l'ordre de 0,5 dans toutes les catégories de fratries.

Parmi les couples de jumeaux, dont il n'a pas été possible de distinguer les monozygotes des dizygotes, les performances apparaissent inférieures à celles des non jumeaux. Les couples de garçons sont inférieurs aux couples de filles. Les coefficients de corrélation sont de 0,842 chez les garçons, 0,823 chez les filles et 0,781 dans les couples hétérogènes en sexe.

British Region, February 26, 1953.

229 KATHERINE H. COWARD. **The Use of the Logistic Curve in Bio-Assay.**

The logistic curve best fitting the data may be determined by strating from trial estimates of the ceiling and base values, and then modifying these values until the average of the ratios of the doses corresponding to successive Y-values most nearly approaches the known ratio of the doses given. Test and Standard can be worked out side by side.

The potency of the Test Substance in terms of Standard is given by (a) the number of units of Standard and (b) the weight of Test Substance corresponding to one standard interval of the curve. (Ref: Emmens, C. W.: J. Endocrinology, 1940, 2, 194).

230 J. O. IRWIN. **On the "Transition Probabilities" Corresponding to Any Accident Distribution.**

From any known distribution of accidents in a fixed exposure time T, the expected number of other accidents sustained by a person who has x accidents is obtained. It is the ratio of the $x + 1$-th to the x-th factorial moment of the distribution. Its limiting value when the exposure time tends to zero gives the transition probabilities. (To appear in J. Roy. Statist. Soc. B.)

231 J. A. FRASER ROBERTS. **The Use of Regressions Involving Variances of Independent Variates for Calculating Age-Corrected Scores.**

It is often useful to calculate once and for all an age-corrected score for each individual. If there is a substantial change of variance with age as well as of mean, the scores can be adjusted by using the regression of the variance of the measurement on age (weighting the variance in each array by the number in the array). The fit of the regression may be tested by a method due to R. A. Fisher.

The mean square, derived by dividing the weighted sum of squares of deviations from the regression line by $k - r$, where k is the number of arrays and $r - 1$ the order of the polynomial is:

$$\sum n_p(y_p - Y_p)^2/(k - r) \qquad (1)$$

The theoretical sampling variance of an estimate y of a true variance Y, based on a sample of n, is $2Y^2/n$. (1) therefore estimates approximately the mean value of $2Y^2$, taken over the k arrays, namely $2 \sum Y_p^2/k$. The approximation, which should lead to little discrepancy, is due to the fact that the observed variances are weighted only according to the number of observations and not according to estimates of their true values. Thus the mean square due to linear regression is compared with twice the square of the mean variance, with degrees of freedom 1 and ∞; the extra mean square due to the quadratic term with $2/k$ times the sum of squares of the expected variances given by the linear function, etc.; the remainder with $2/k$ times the sum of squares of the expected variances given by the highest polynomial fitted, with degrees of freedom $(k - r)$ and ∞.

Application of the method to the Terman-Merrill revision of the Binet Scale removes the great bulk of the very large and troublesome residual association between age and I.Q.

Mean blood-pressure grows steadily throughout life and variance increases more than fivefold. In a genetic investigation it is essential to be able to compare and add in various ways persons of very different ages, for example, the parents, the sibs and the children of a sample of subjects. Adjustment for variance as well as for mean has provided individual scores which are almost entirely independent of age. (See J. A. F. Roberts and M. A. Mellone. British Journal of Psychology, Statistical Section, *5*, 65, 1952; G. W. Pickering, G. S. C. Sowry, M. Hamilton and J. A. F. Roberts. Clinical Science, in preparation.)

THE THIRD INTERNATIONAL BIOMETRIC CONFERENCE

Preliminary Programme

Lake of Como, Bellagio, Italy—September 1-5, 1953

1st September

9:00 A.M. Inauguration: Welcoming address and Presidential address.

10:00 A.M. *The first course in biometry—a symposium* (Chairman: W. G. Cochran)

L. Martin. Enseignement des principes d'expérimentation des méthodes statistiques à des biologistes dans deux établissements belges d'enseignement supérieur.

G. Barbensi. L'insegnamento della Biometria.

C. I. Bliss. A course in Biometry for graduate students in Biology.

By Title: A. Vessereau. Enseignement des methodes statistiques appliquées à la Biometrie.

12:30 P.M. First general business meeting.

3:00 P.M. *Mathematical problems in Genetics* (Chairman: A. Buzzati-Traverso)

Sir Ronald Fisher. The variability in the length of germ plasm still heterogeneous after a given amount of inbreeding.

K. Mather. The methodology of Biometrical Genetics.

A. R. G. Owen. Experimental designs in Genetics.

D. Lowry. Variance components with reference to genetic population parameters.

C. A. B. Smith. The calculation of correlation between cousins.

2nd September.

9:00 A.M. *Methodological Problems in Biometry* (Chairman: Gertrude M. Cox)

J. W. Hopkins. Some needed significance tests.

F. Anscombe. Fixed-sample-size analysis of sequential observations.

W. G. Cochran. The combination of estimates from different experiments.

N. Blomqvist. Rank analysis of incomplete block designs.

M. Keuls. Testing differences between means in an analysis of variance.

M. J. R. Healy. Decision between two alternatives: how many experiments?

By Title: L. Martin. Suggestions for longitudinal data in gerontology.

J. Hemelrijk and J. H. B. Kemperman. Use of a sequential Wilcoxon-test for diagnostic purposes.

3:00 P.M. *Biometry in Immunology* (Chairman: H. C. Batson)

R. Prigge. Die Anwendung der Mutungsbereiche in der Immunitaetsforschung.

I. Ipsen. Factors of dosage and host determining antibody response to secondary antigen stimulus.

L. B. Holt. Quantitative studies in diphtheria prophylaxis—an attempt to derive a mathematical characterisation of the antigenicity of diphtheria prophylactics.

S. Peto. A dose response equation for the invasion of microorganisms.

3rd September.

9:00 A.M. *Biometric methods in Agriculture* (Chairman: P. V. Sukhatme)

F. Yates. The place of simple experiments on cultivator's fields in agricultural development.

V. G. Panse. Principles of the survey method of experimentation.

T. N. Hoblyn and S. C. Pearce. Biometrical problems in research on long-lived plants.

H. Strecker and J. Raab. Methodological problems in a survey of milk production of hog breeders in Germany.

G. Rasch. On different sources of errors and the advantage of their knowledge in planning experiments.

J. L. Lush. Estimating heritabilities.

M. Matemura. Designs for agricultural research in Japan.

2:30 P.M. Excursion.

4th September.

9:00 A.M. *Functional relations in experimentation* (Chairman: H. Wold)

D. J. Finney. Functional relationships in experimentation.

J. Berkson. Minimum chi square and maximum likelihood estimates of regression coefficients.

J. Neyman. Frisch's problem on linear structural relations.

12:00 A.M. Biometrics business meeting.

3:00 P.M. *Contributed papers* (Chairman: M. J. R. Healy)

A. F. Parker-Rhodes. Estimating populations of irregularly observable organisms.

D. W. Goodall. Factor analysis in Plant Sociology.

E. F. Scott. Bivariate contagious distributions.

G. Karreman. The mathematical biology threshold and related phenomena in excitation.

By Title: M. W. Bentzon. On the statistical evaluation of dose response curves in case the dose intervals are large.

8:00 P.M. Social dinner.

5th September.

9:00 A.M. *Industrial applications of Biometry* (Chairman: A. Linder)

E. A. G. Knowles. Applications of experimental designs in industry.

D. R. Read. The design of chemical experiments.

H. C. Hamaker. Experimental designs in industry: a discussion.

11:30 A.M. General business session.

12:00 A.M. Closing of the Conference.

THE BIOMETRIC SOCIETY

Indian Members. The Indian group sponsored jointly with the Indian Society of Agricultural Statistics a symposium in New Delhi on the 26th of February on "New Problems in Experimentation." Frank Yates presided and the speakers included D. J. Finney, K. R. Nair, V. G. Panse and others.

British Region. The British Region held a joint symposium at the Wellcome Research Institution in London on March 11th on "Flavour Assessment" in collaboration with the Society of Chemical Industry (Food Group) and the Society of Public Analysts (Biological Methods Group). The program consisted of papers by E. D. Adrian on The physiological background of flavour assessment, by H. G. Harvey on Basic considerations in regard to flavour assessment, by A. S. C. Ehrenburg and J. M. Shewan on The objective approach to sensory tests, by J. M. Harries on Sensory tests and consumer acceptance, and by J. O. Irwin on A biometrician's viewpoint. A more complete account of the meeting has been published in *Nature* for April 25, 1953.

Italian Region. The Third Italian Meeting was held at the University of Florence on April 2nd. In the morning session on "Genetic Selection" papers were presented by I. M. Lerner, currently a guest at the University of Pavia, on The problem of quantification of selection methods, and by R. Scossiroli, also of Pavia, on Statistical analysis of selection experiments in *Drosophila* populations treated with X-rays. The afternoon program on "General Methodology" offered lectures by F. Brambilla, of the University of Genoa and the University Bacconi, Milan, on Confluence analysis, and by G. Barbensi on Criticisms on the use of chi-square in the fourfold table. At the annual business meeting, which followed the morning scientific program, L. L. Cavalli reported on the activities of the year. Statutes for the Region were discussed and approved. It was agreed to send a letter over the signatures of several university professors in medical and biological faculties recommending that these faculties provide teaching in statistics. This letter is to be given as wide circulation as possible. The Region will sponsor a duplicated bulletin of methodological papers of general interest, with emphasis on the popularization of statistical methods. It will be sent without charge to each member. Through the initiative of Professor Brambilla, a Methodological Section has been formed which holds weekly seminars at the Department of Statistics of the University Bacconi of Milan.

ENAR. On April 29th to May 1st the Eastern North American Region met jointly with the Institute of Mathematical Statistics in Washington, D. C. The opening session on April 29th considered "Statistics in the Physical Sciences" with papers by J. E. Jackson and R. H. Morris on The application of multivariate quality control to a photographic problem, and by J. M. Cameron on Control and measurement of experimental error. On Thursday "Statistics in the Biological Sciences" provided the morning program, with papers by H. C. Batson on Factorial chi-square analysis of data from experiments in immunology, by I. D. J. Bross on Applications of nonparametric methods to medical data, by D. G. Kendall on Stochastic growth and mutation processes, and by G. E. P. Box and W. A. Hay on Statistical designs for the efficient removal of trends occurring in comparative experiments with applications in biological assay. The afternoon program offered papers by W. H. Clatworthy and W. S. Connor on Necessary conditions for the existence of partially balanced incomplete block designs with two associate classes, by A. C. Cohen, Jr. on Estimation in truncated bivariate normal distributions, by R. F. Drenick and P. Nesbeda on A class of optimum linear predictors, by D. B. Duncan on Multiple range tests and the multiple comparisons test, by N. L. Johnson on Sequential procedures in component of variance problems, and by M. C. K. Tweedie on Sequential estimation. On Friday papers were given by M. Zelen on The analysis of some incomplete block designs with a missing block, by A. W. Kimball on The fitting of multi-hit survival curves, by C. I. Bliss, T. Greiner and H. Gold on Estimating the dose of a cardiac glycoside for human subjects, by C. Y. Kramer On the analysis of variance of a two-way classification with unequal sub-class numbers, by A. B. Parks on Ranking versus scoring in palatability tests using small, trained panels, by M. A. Bershad, W. N. Hurwitz and R. S. Woodruff on Sampling for time series, by W. A. Hendricks on Response rates and selectivity in mail surveys, by R. C. Bose and S. N. Roy on Simultaneous confidence interval estimation and testing of hypotheses.

Japanese Members. Largely through the efforts of Mr. M. Hatamura of the National Institute of Agricultural Sciences in Tokyo, the Society has at this date 25 members in Japan. Mr. Hatamura has agreed to serve provisionally as National Secretary. He reports that a meeting is planned in the near future to develop a program for Japan.

NOTES

Summer Statistical Seminar

University of Connecticut, 1953

The Summer Seminar in Statistics will meet for the fourth year at the University of Connecticut during the three weeks of August 10 through 28, 1953. The general pattern of programs is to be the same as in previous years. Informality and discussion will be stressed. There will be one or two seminar sessions each day and a clinic on the treatment of problems in application.

The first week, August 10 through 14 will be devoted to Statistical Methodology in Physics. It is being organized by Dr. E. W. Pike, Equipment Engineering Division, Raytheon Manufacturing Company, Newton 58, Mass., together with Dr. Churchill Eisenhart and Dr. E. P. King.

The second week, August 17 through 21 will be devoted to Statistics in Biometry and Medicine. It is being organized by Professor G. Beall, Statistical Laboratory, University of Connecticut, Storrs, Conn., together with Dr. I. Bross and Dr. D. Mainland.

The third week, August 24 through 28 will be broken into two parts. The first will be devoted to the ASA Handbook, as organized by Professor F. Mosteller, Laboratory of Social Relations, Harvard University, Cambridge 38, Mass. The second part will be devoted to Performance and Reliability of Complex Mechanical Assemblies, as organized by Professor G. H. Shortley, The Johns Hopkins University, 6410 Connecticut Avenue, Chevy Chase, Maryland.

Anyone interested in the subjects under discussion is invited to attend for the day, week, or other period. (A nominal registration fee will be collected).

For further discussion please write the Secretary of the Seminar, Professor Geoffrey Beall, Statistical Laboratory, University of Connecticut, Storrs, Connecticut. Information will be provided on lodging and meals. A detailed outline of the program will be furnished. In addition to supplying information, the Secretary will welcome suggestions of topics for the Seminar or Clinic sessions.

International Congress of Mathematicians 1954

First Communication

In the final plenary session of the International Congress of Mathematicians, 1950, held in Cambridge (Mass.) the Congress accepted the invitation of the delegation from the Netherlands to hold the next Congress in the Netherlands.

The International Congress of Mathematicians 1954 will be held in Amsterdam from September 2nd to September 9th under the auspices of "Het Wiskundig Genootschap" (The Mathematical Society of the Netherlands). It is the sincere hope of the "Wiskundig Genootschap" that the Congress 1954, which will be open to all mathematicians from all parts of the world, will be a fertile international gathering.

The Organizing Committee has invited a number of outstanding mathematicians to deliver one-hour addresses, hoping that in this way a survey of the recent development in the whole field of mathematics may be furnished.

There will be seven sections, viz: (1) Algebra and Theory of Numbers, (2) Analysis, (3) Geometry and Topology, (4) Probability and Statistics, (5) Mathematical Physics and Applied Mathematics, (6) Logic and Foundations, and (7) Philosophy, History and Education.